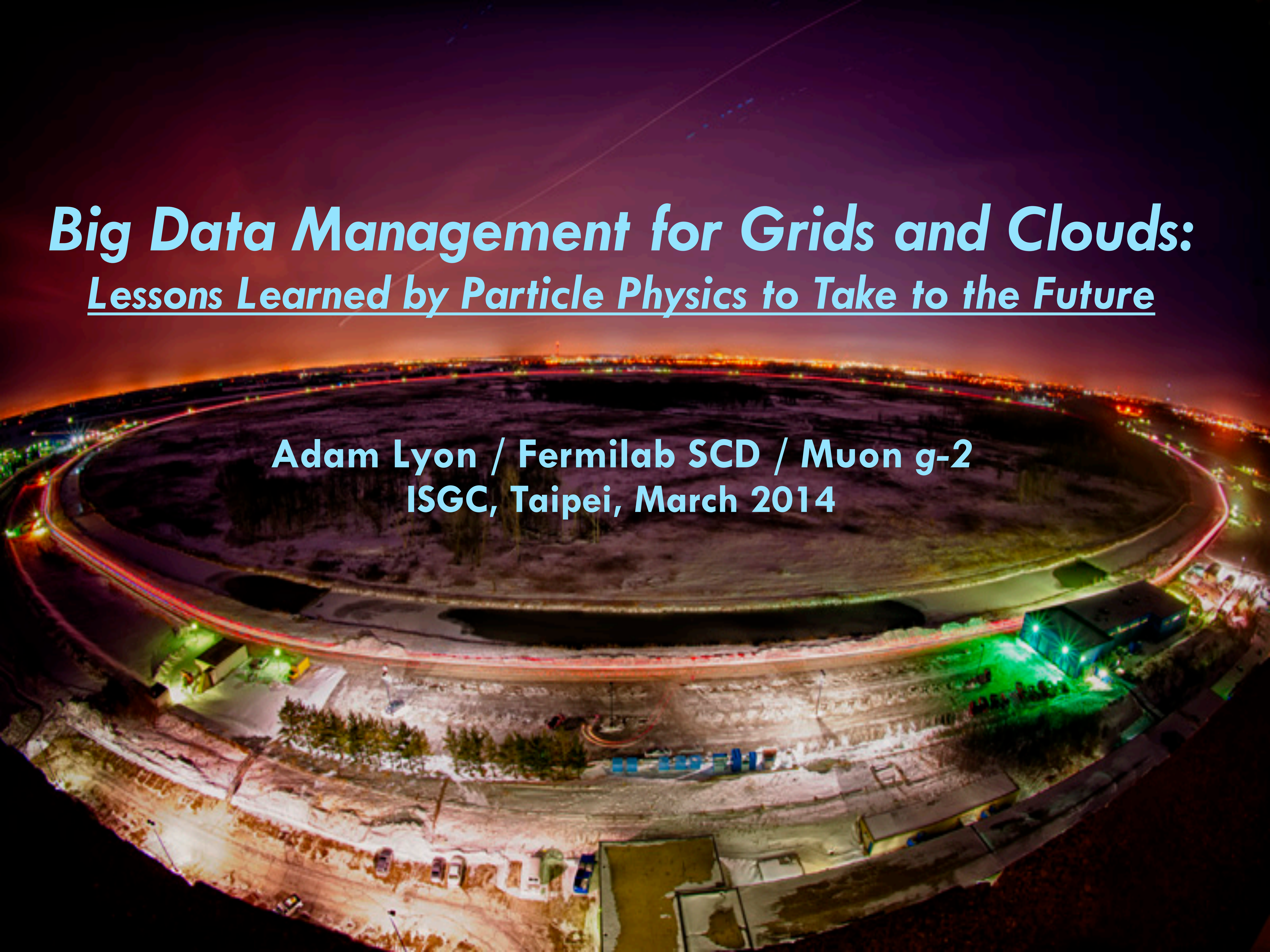


Big Data Management for Grids and Clouds:

Lessons Learned by Particle Physics to Take to the Future

Adam Lyon / Fermilab SCD / Muon *g*-2
ISGC, Taipei, March 2014

Some context

Experimental Particle Physics in 2 minutes…

Collisions

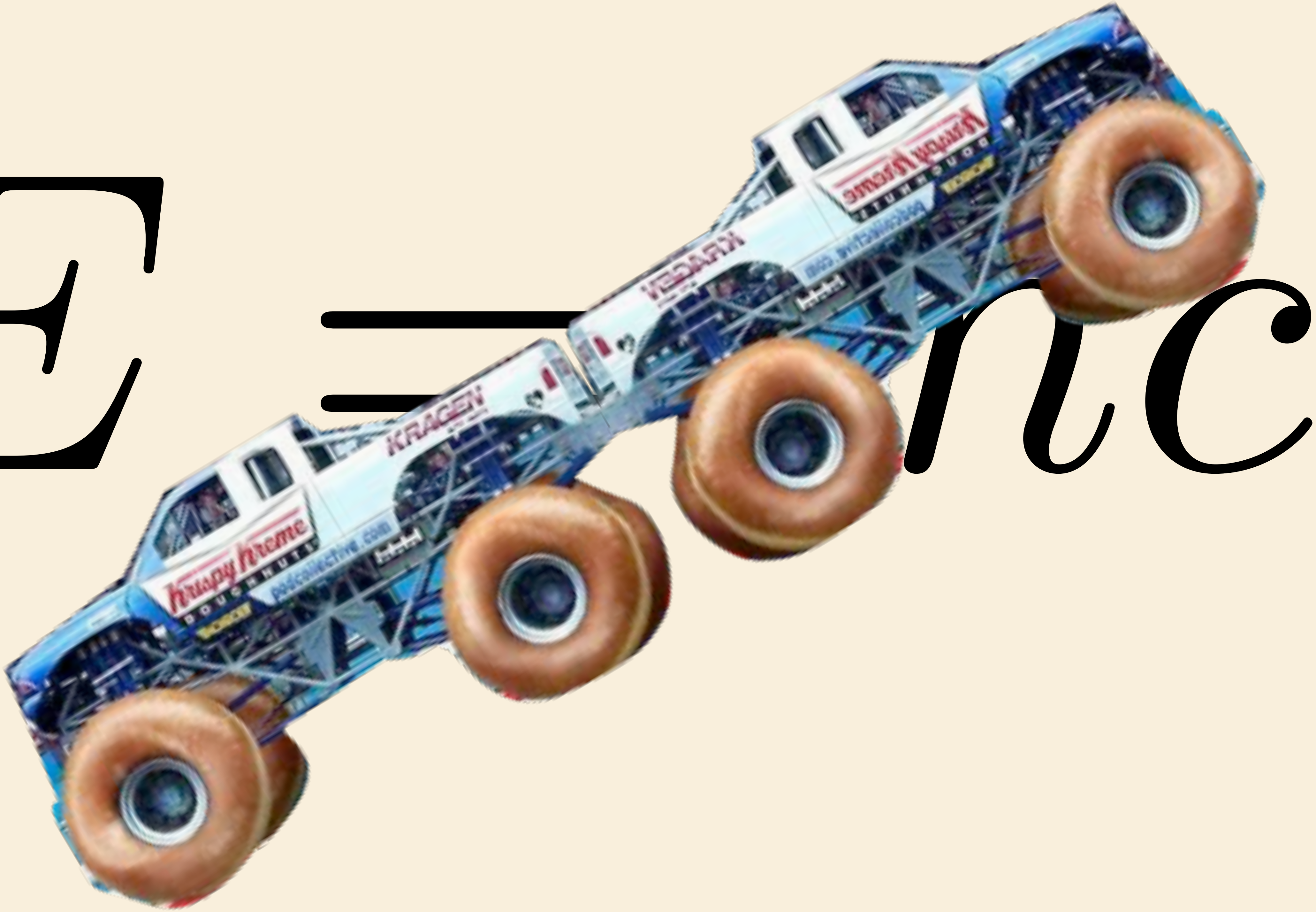
$$E = mc^2$$

Collisions

$$E = mc^2$$

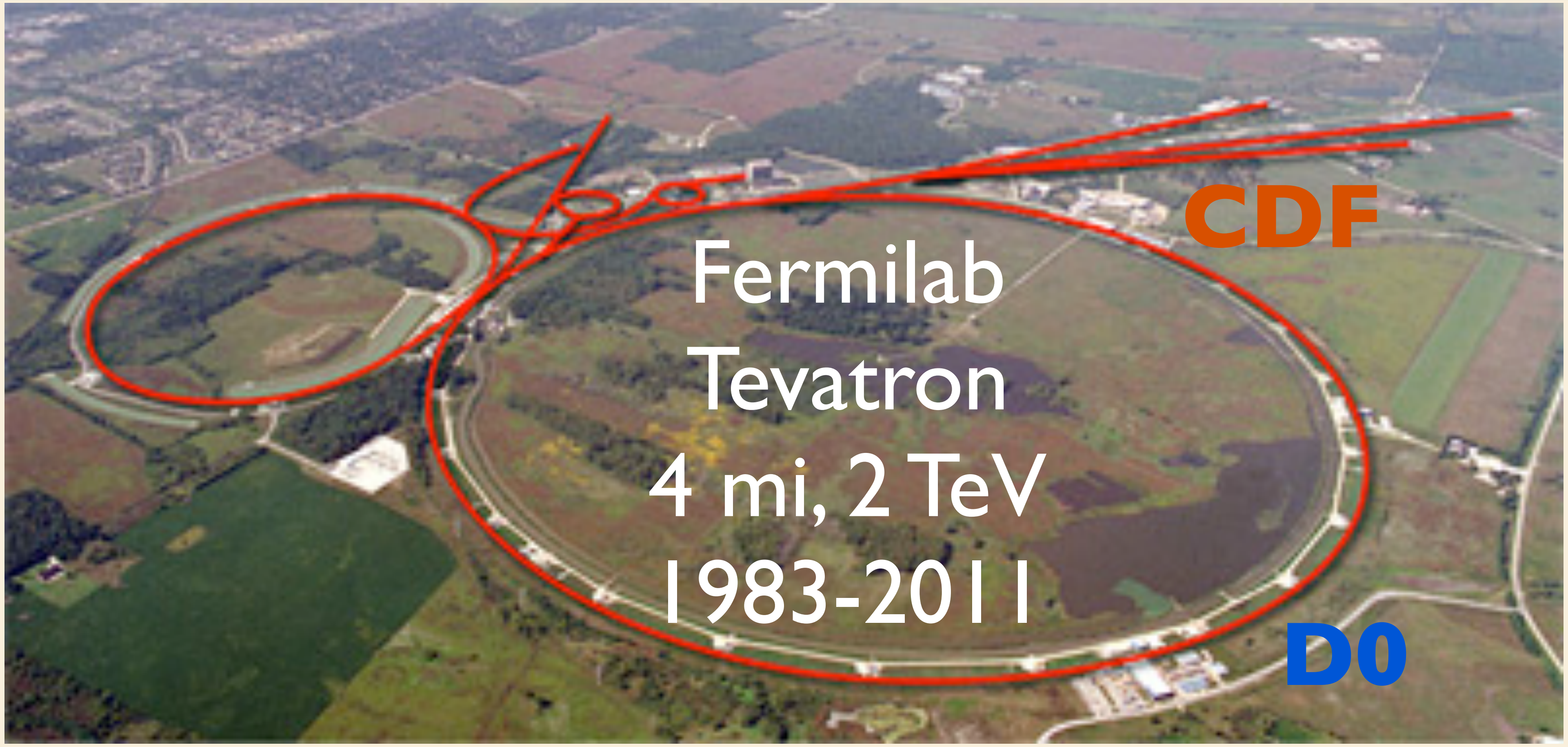
Collisions

$$E = mc^2$$



Collisions

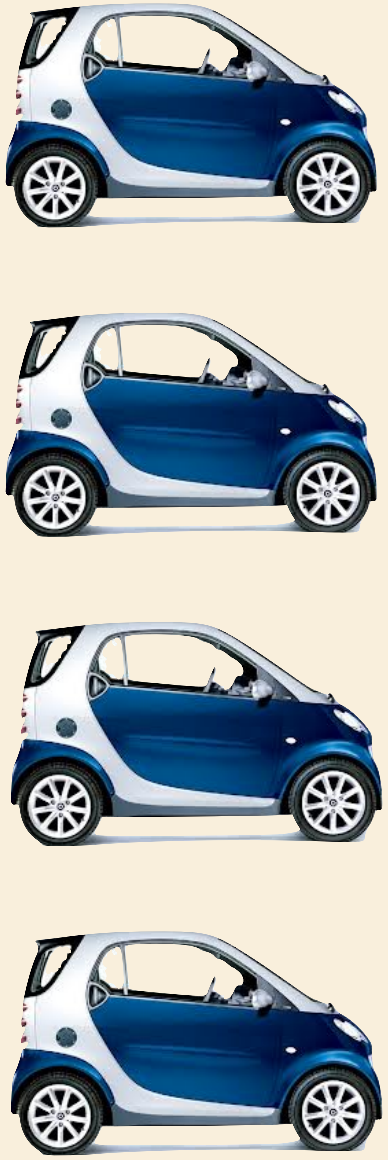
The Energy Frontier



Fermilab
Tevatron
4 mi, 2 TeV
1983-2011

CERN
Large Hadron Collider
17 mi, 7–14 TeV
2010–

Neutrino Oscillations



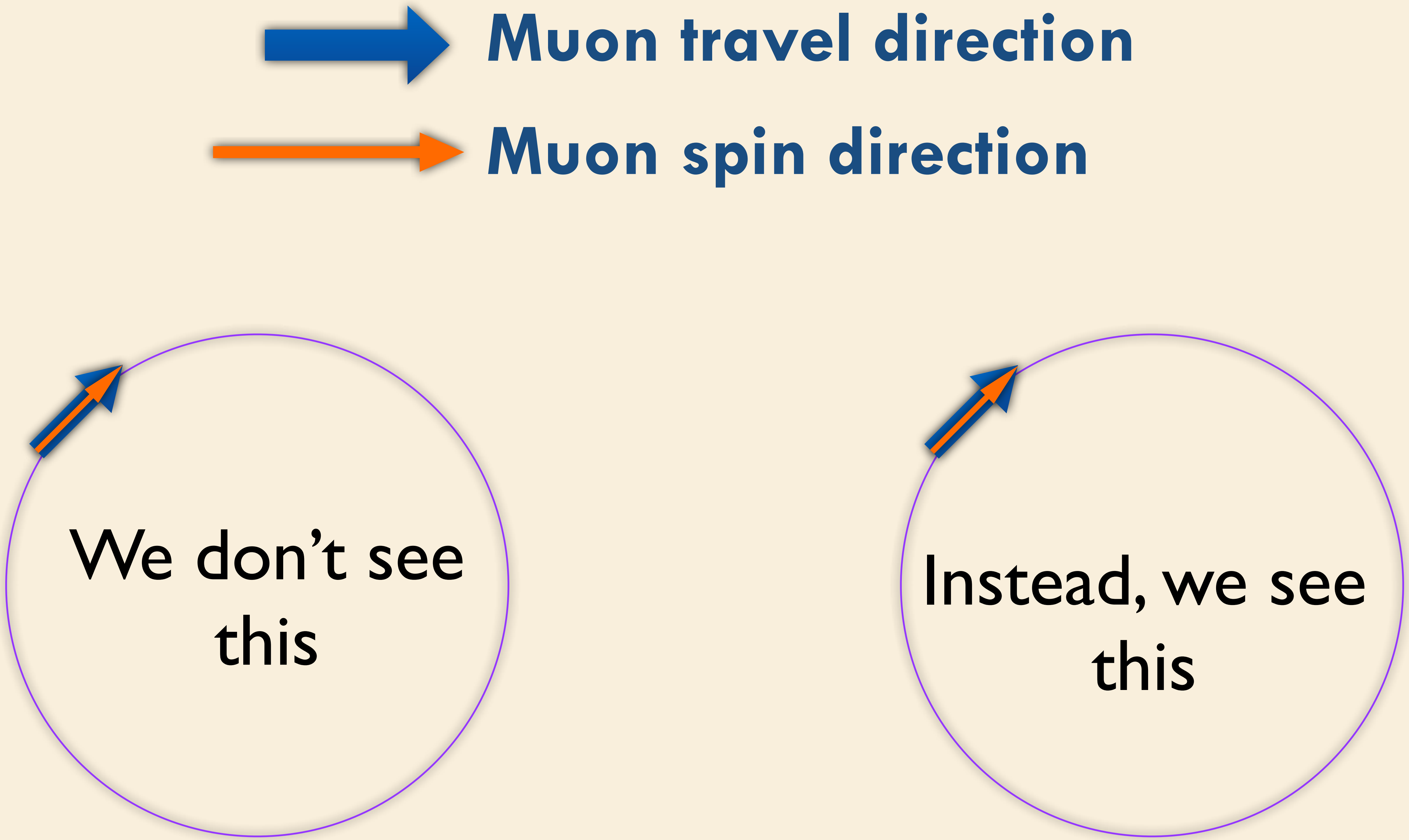
Neutrino Oscillations

Neutrino Oscillations

NOvA, MicroBoone, LBNE
Dya Bay, T2K, OPERA

Muon g-2

Muon in a magnetic storage ring



Muon travel direction

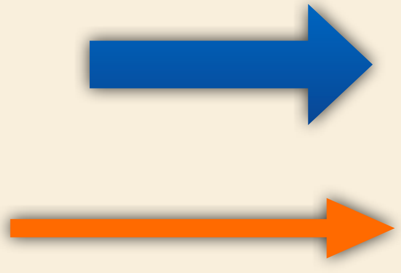
Muon spin direction

We don’t see this

Instead, we see this

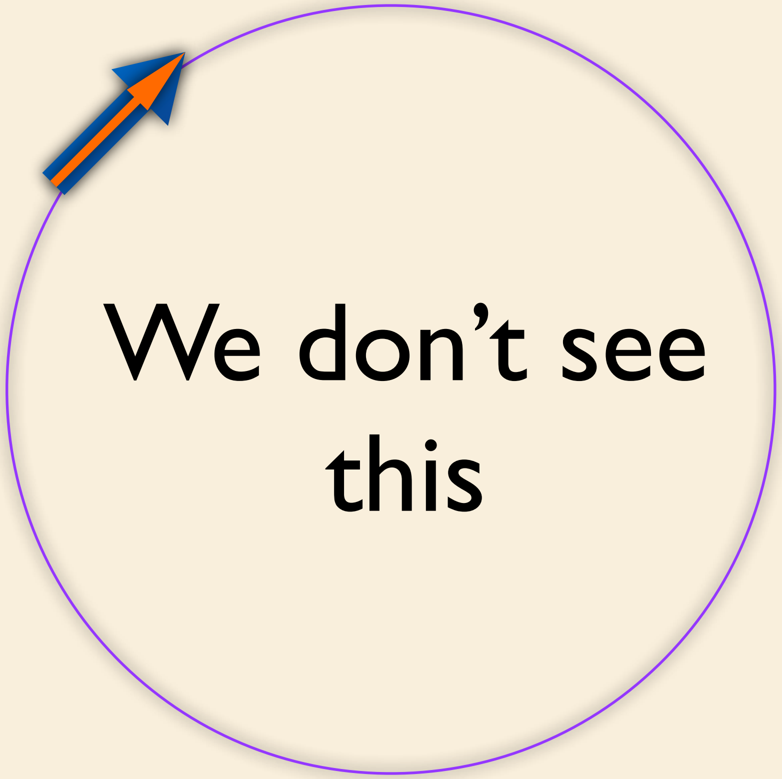
Muon g-2

Muon in a magnetic storage ring

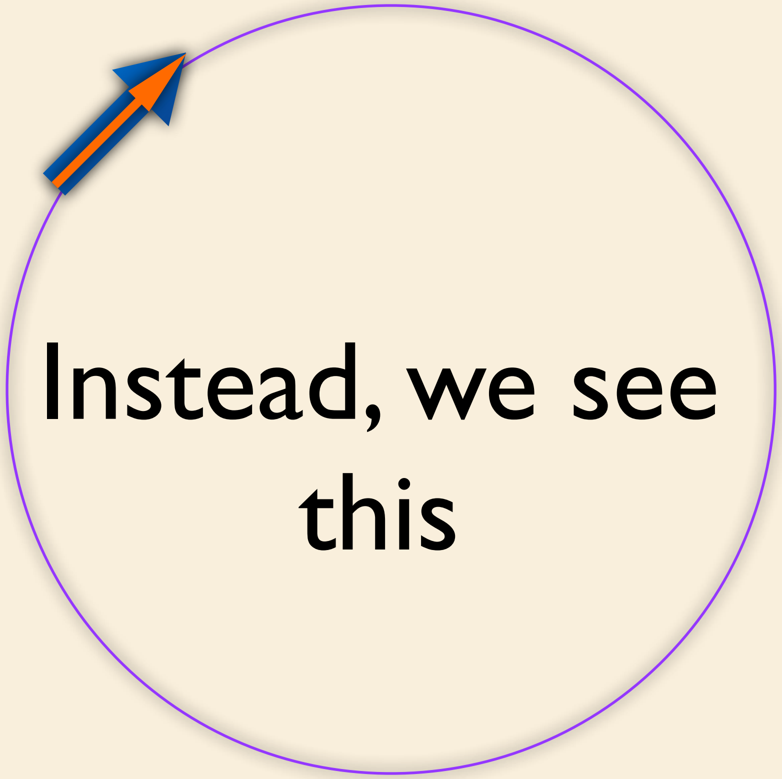


Muon travel direction

Muon spin direction

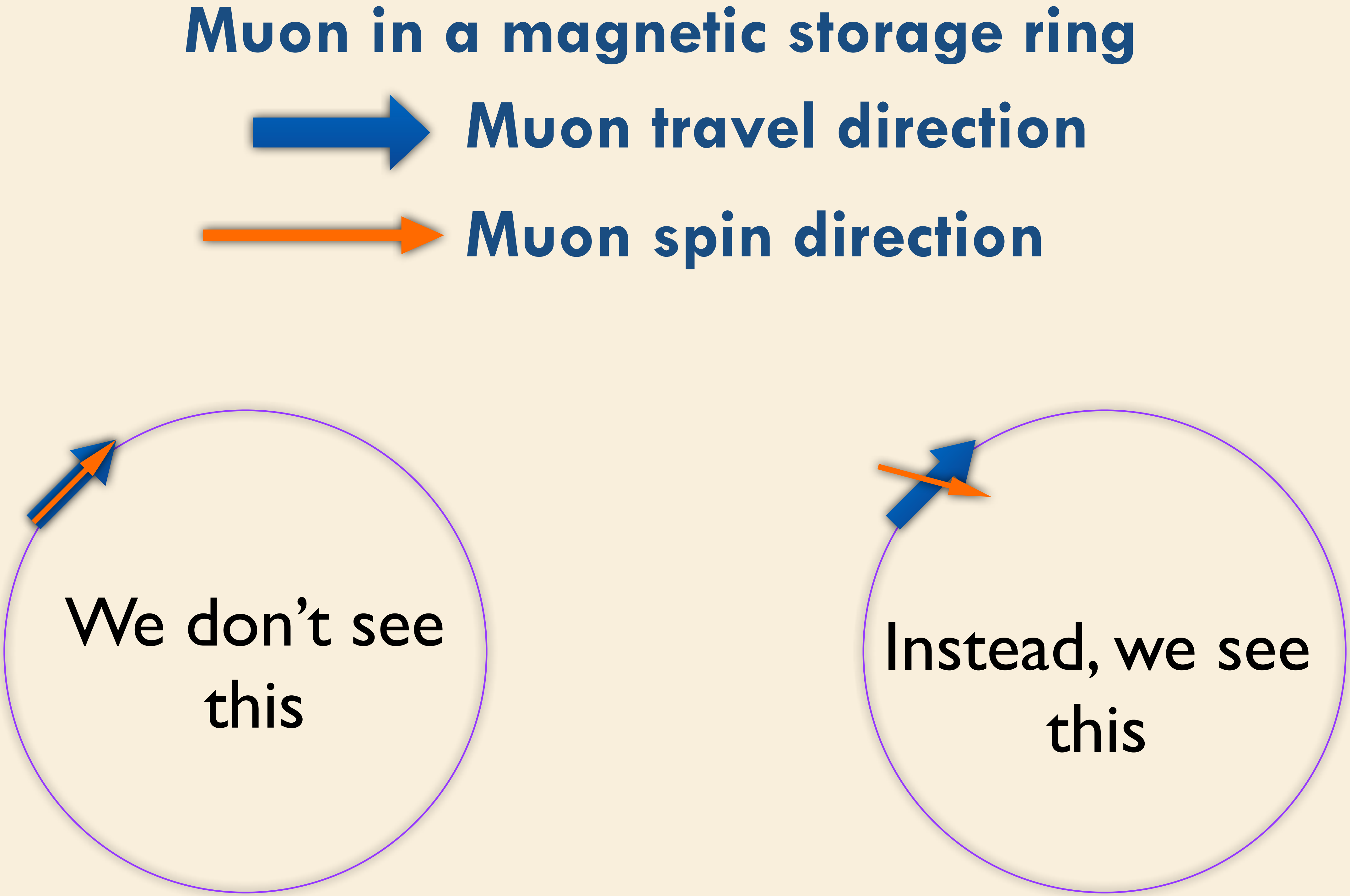


We don't see this



Instead, we see this

Muon g-2



Muon in a magnetic storage ring

Muon travel direction

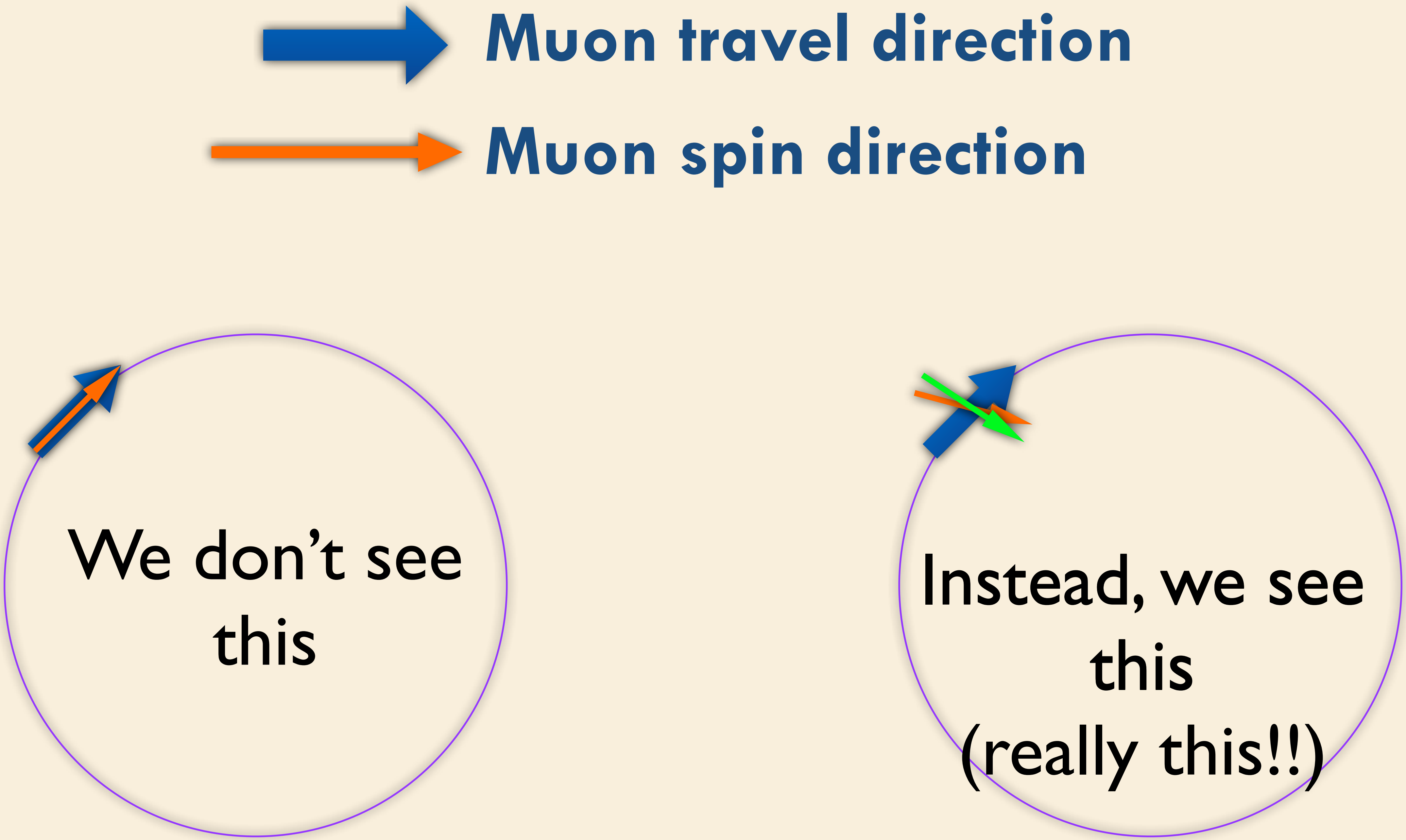
Muon spin direction

We don't see this

Instead, we see this

Muon g-2

Muon in a magnetic storage ring



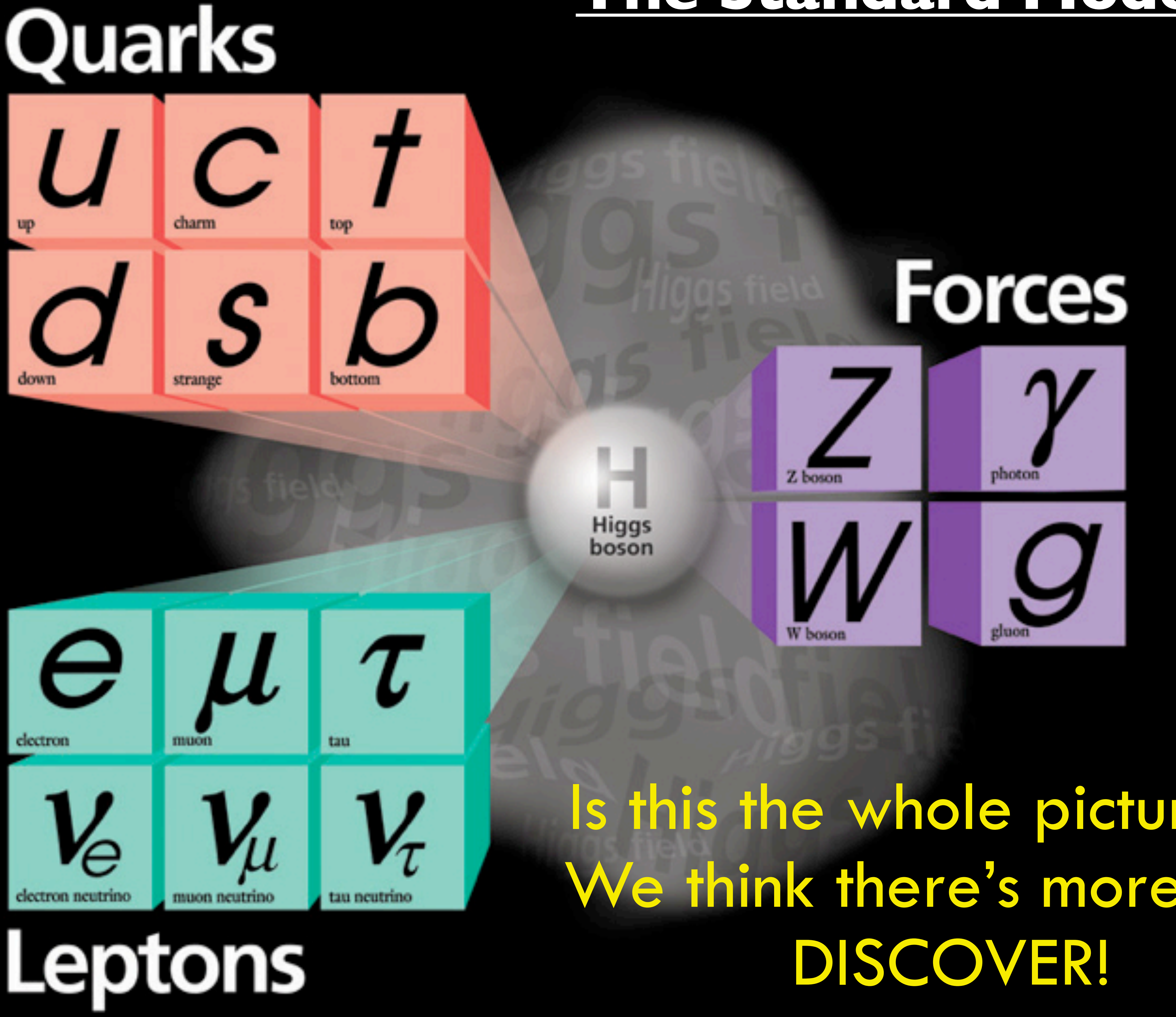
Muon travel direction

Muon spin direction

We don't see this

Instead, we see this (really this!!)

The Standard Model

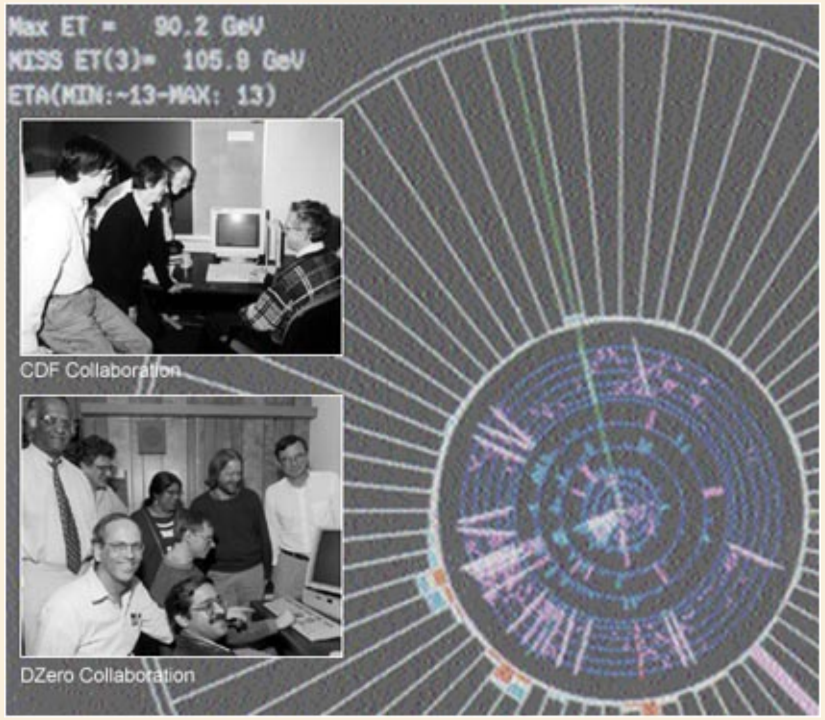


Is this the whole picture?
We think there's more to
DISCOVER!

The next level of discoveries…

The next level of discoveries requires the next level of…

W boson 1983

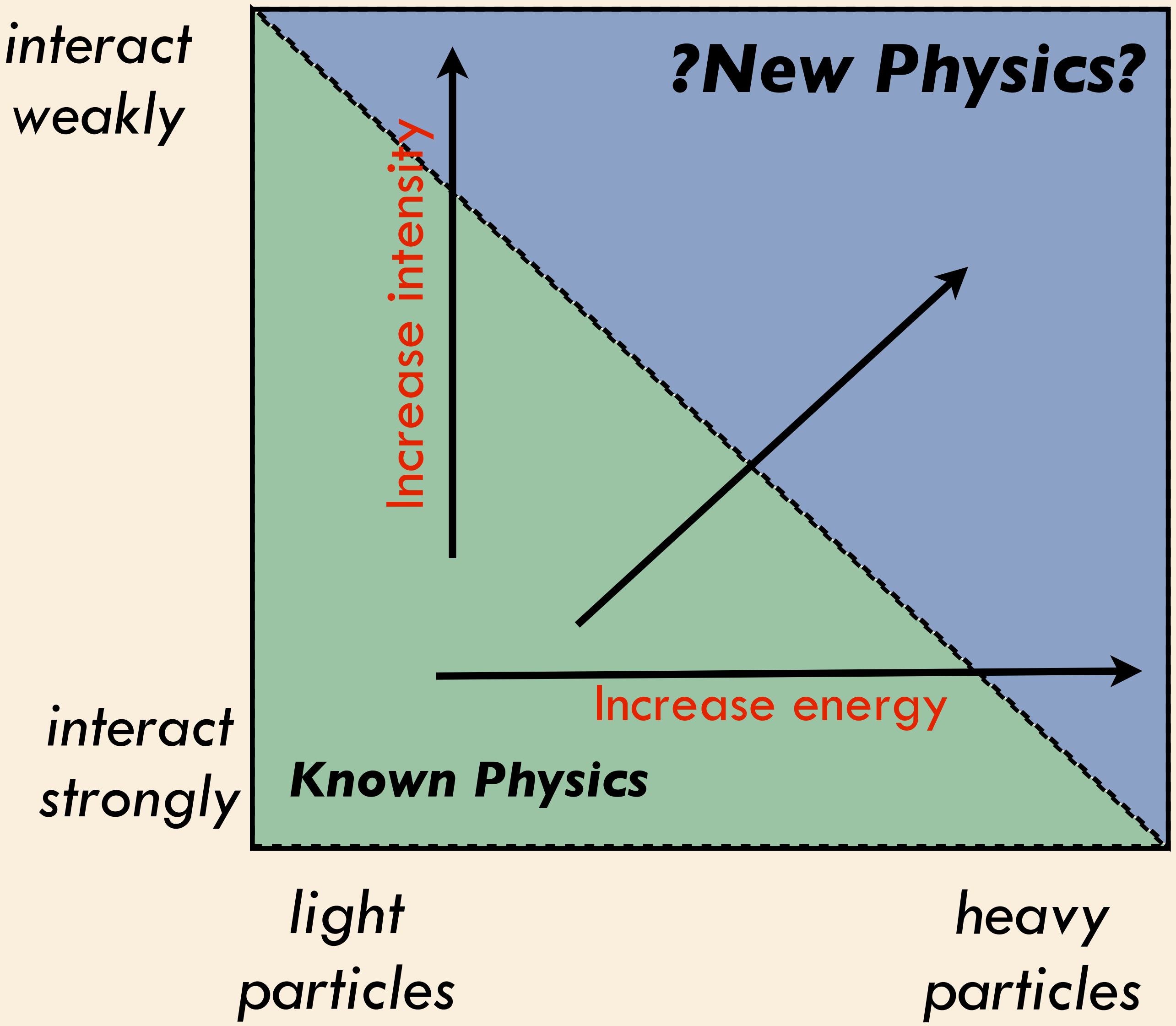


Top Quark 1995

Higgs Boson 2012

… the next level of …

New physics hides behind particles too heavy to make or too weakly interacting



The next level of Energy
The next level of Intensity
The next level of Analysis

along with these come…

The next level of Accelerators
The next level of Detectors
The next level of Computing

The next level of technology
The next level of collaboration

The next level in context of big data

Let's study the evolution of data management and big data

The challenge of Data Management has had many solutions

Examine lessons learned from past and current systems

How can they guide the future of computing, data management and Big Data?

The following is a personal tour.

Allons-y!

The basic problem

Your files reside here

Tape Robot

Disk Storage

Not here, where you need them

Compute Farm of worker nodes

Must identify the files to process

Must move those files to the right place

Must associate those files with jobs and process them

All this must be reliable, scalable, and efficient

My story

‘02- Staff Scientist at Fermilab in Scientific Computing Division – Head of “Scientific Data Processing” Department, Worked on Tevatron Run 2 D0 and now Muon *g-2*

Was project manager for “SAM” (Data Management system for Fermilab Tevatron Run II experiments and now Intensity Frontier)

‘97-02 U Rochester Postdoc on CLEO II.5/III @ Cornell ($b \rightarrow s\gamma$)

‘91-97 U Maryland Grad Student on Tevatron Run 1 D0 @ Fermilab (Supersymmetry)

The Particle Physics Data Management Story

Why does particle physics push the data management envelope?

The data volume is ever increasing and increasing complexity of detectors

scalability

CDF & D0 Run II: ~1M readout channels

Atlas & CMS: ~100M channels

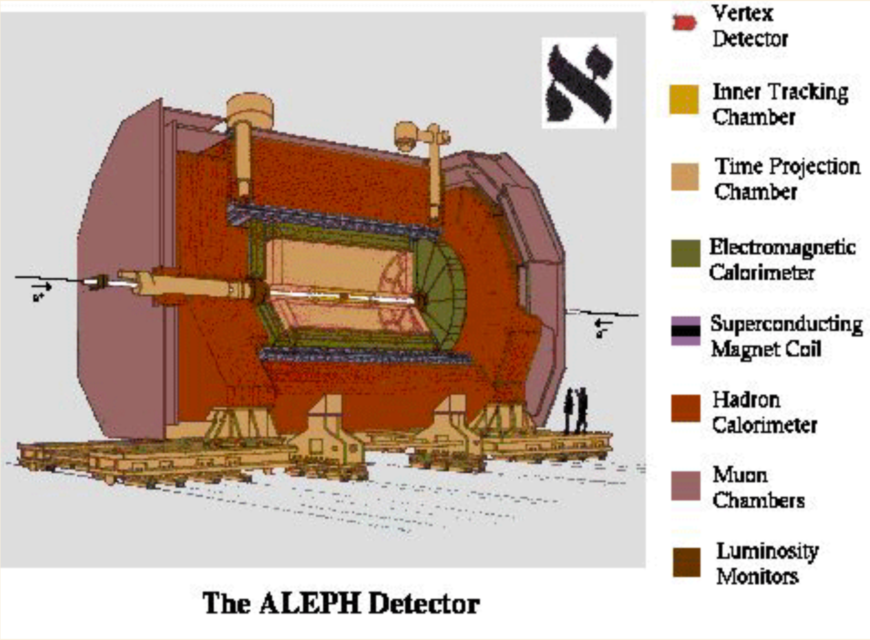
Increasing number of collisions (bigger haystack to find more needles)

CDF & D0: ~ 10 fb^{-1} (~8 PB) in 10 years

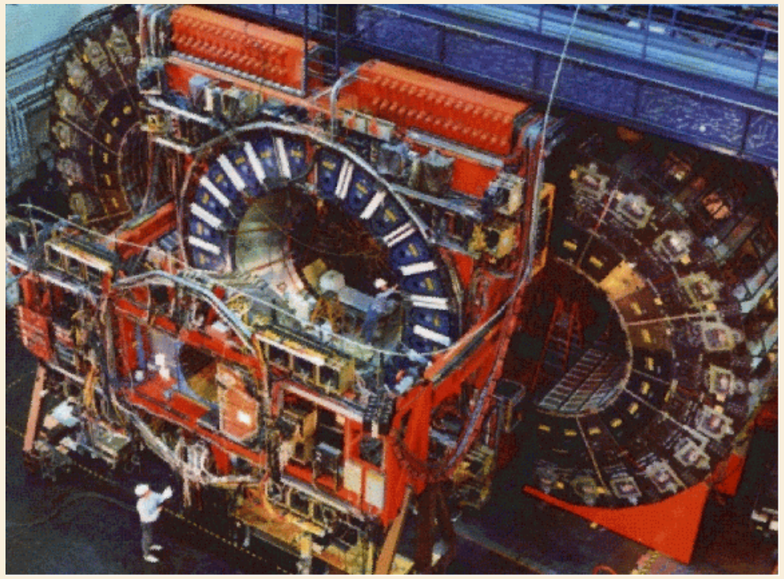
Atlas & CMS: ~25-30 fb^{-1} (~150 PB) in 3 years

e.g. Evolution in size (note people scale)

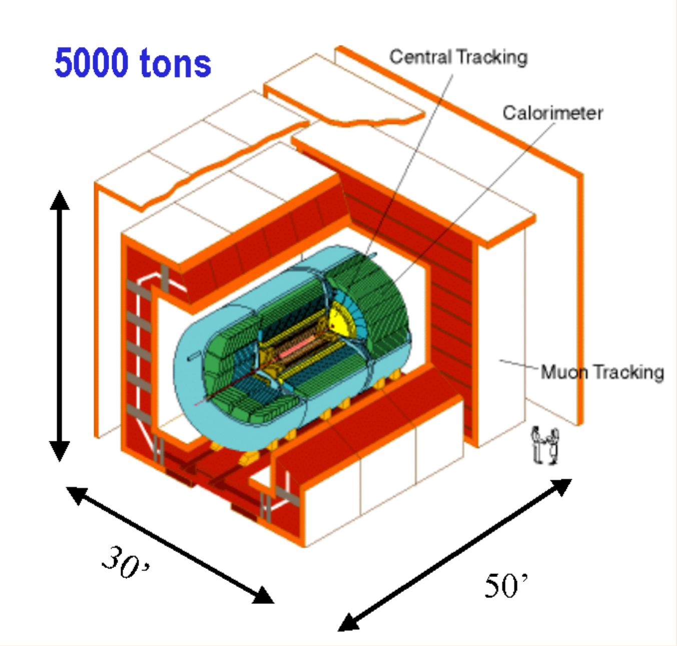
Aleph '89–'00



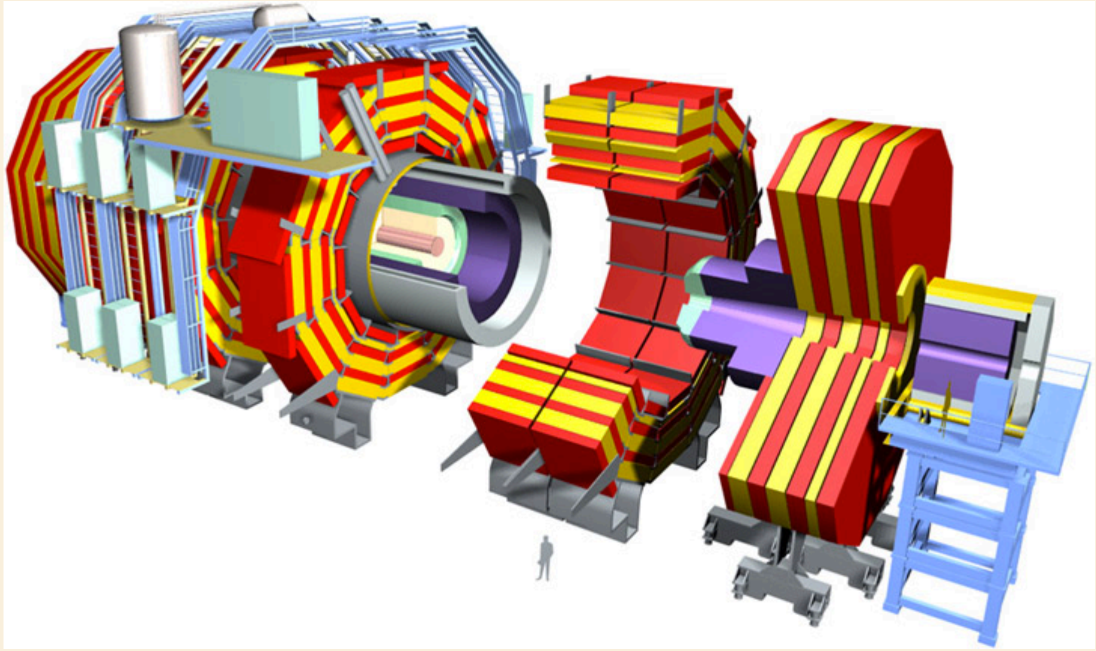
CDF '85–'11



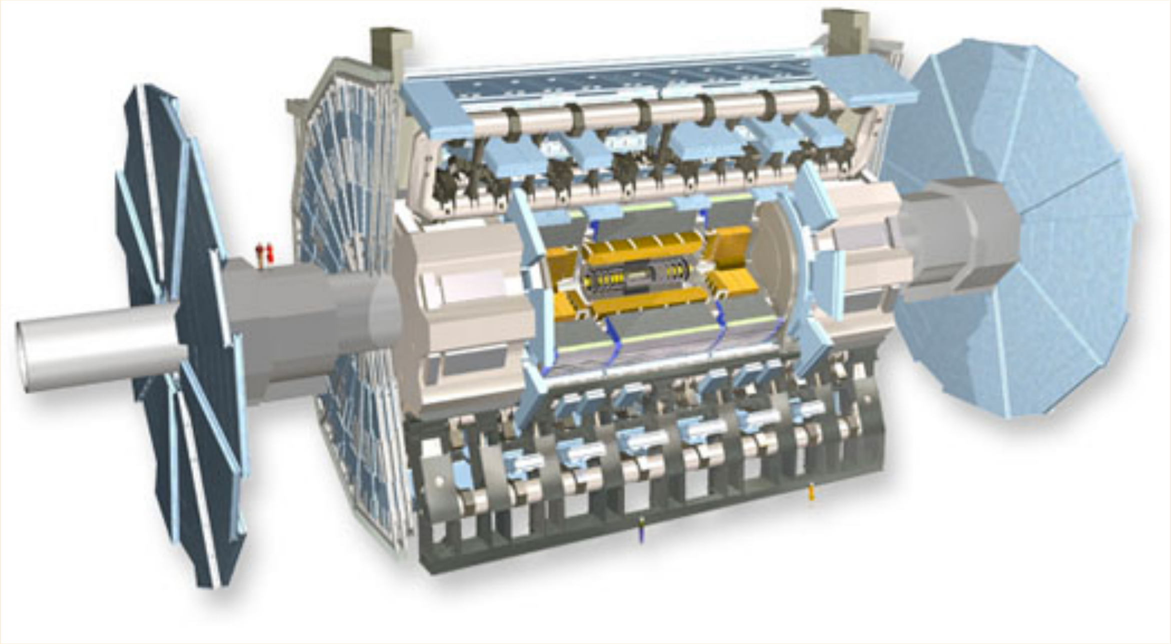
D0 '92–'11



CMS '10–



Atlas '10–



Pre-Grid Data Management

- My experience: Run 1 D0 ('93-97)
- FORTRAN, Zebra Banks, HBOOK, PAW
- Most processing on laboratory clusters (VAX, SGI, IBM RS-6000, DEC Alpha)
- 8mm tapes, mounted by human operators
  (evolved to small robots, then HPSS)
  Now use T10000 tapes @ FNAL
- Some Monte Carlo generation offsite, but not centrally organized
- Tapes from remote sites came to Fermilab with traveling students/professors or shipped in mail
- Some (lucky) students processed data using sizable resources at their university

reliability

efficiency

Pre-grid lessons

- Don't underestimate the bandwidth of transporting data by truck (or suitcase) [efficient, but not scalable]
  Modern equivalent is AWS import/export service: (ship your hard drives to Amazon and they'll upload the data to their storage)

- Offsite processing was an interesting challenge, often not centrally organized, offsite resources for particular groups

- By today's standards not very reliable, scalable, or efficient
  But good enough to get the job done (discovered top quark)

Data Type	DØ Run 1	CDF Run 1
Raw	30 TB, 500 KB/ev	8 TB, 130 KB/ev
Derived	25 TB	33 TB
TOTAL	55 TB, 42M events	41 TB, 64M events

A paradigm shift: transition to C++/OO

My experience (D0 and CLEO ’97-02)

FORTRAN did not have the data organization necessary for the next level of data complexity

Industry and education shifting away from FORTRAN

Object oriented programming and C++ were adopted by cadres of students and postdocs

Lesson Learned: Large scale C++ systems are not easy to write.

Software beginning to become a main part of the experiment, like the detector hardware
(needs extensive planning, design, construction, upgrades)

An excursion away from the file... OO Databases!

Objectivity Database

- Persistency Mechanism:
  The form of data on storage (not same as in memory, especially due to pointers)
  Translation is overhead - must think hard to get right (efficiency)

- Several experiments adopted Objectivity (e.g. BaBar, Cleo III)
  - Object Oriented DB Management System
  - Automated persistency - retained object structure
  - Retrieve events with queries
  - Hot (disk) and cold (tape) stores
  - A great idea!
  - In 2003 BaBar held the world record for the largest civilian hybrid (disk & tape) database - 1 PB

But... scalability

For event databases, Objectivity didn't scale:

- Parallel access problematic (hint of things to come)
- $$$
- Large data volume overhead
- Hard to distribute the data
- Federation system difficult
- Reliance on commercial proprietary format

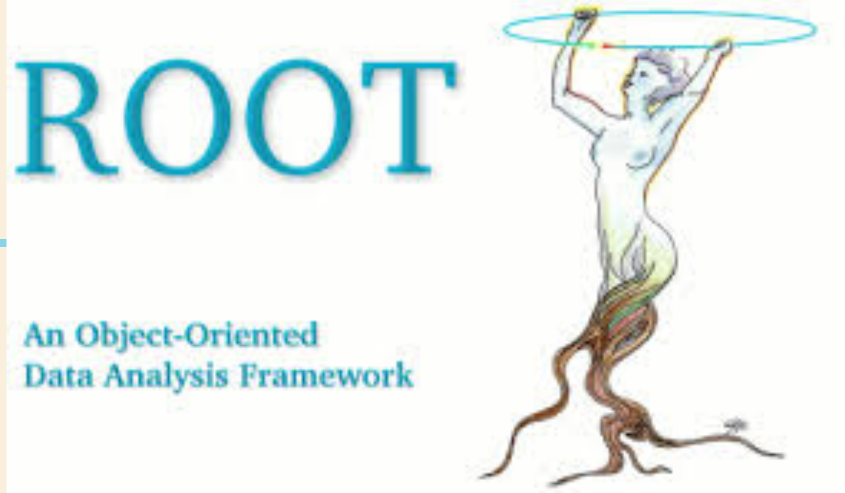
Most experiments returned to the file paradigm by 2004-05 (BaBar – Root; CLEO – PDS [home grown])

Lesson:

A storage system is only as great as how it scales !

An important lesson. Objectivity was an interesting experiment.

The rise of ROOT

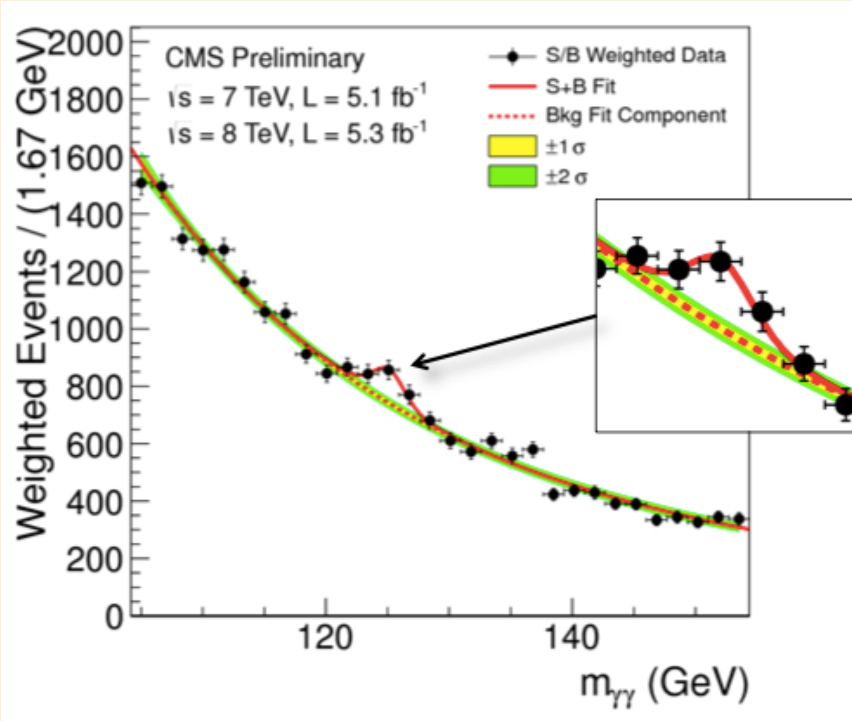


- Analysis platform and toolkit
- Root is to C++ like PAW is to FORTRAN/HBOOK
- In '97 I saw first Root demo at Fermilab (reliability)

- Embraced data as C++ objects – helped to solve persistency problem

- Love it or not, it is ubiquitous in Experimental Particle Physics

- My opinion: Dedication of developers to user base was key to success – continues



- Root I/O and files are also ubiquitous and have evolved. Having a common format leads to efficiency (e.g. XRootD). Very important to the field.

- Lesson: Paradigm shifts require huge commitments.

  Root survives by always keeping the next level in sight

Data Handling for Tevatron Run II

A next level

1997 – Joint CDF/DØ Data Management Needs Assessment

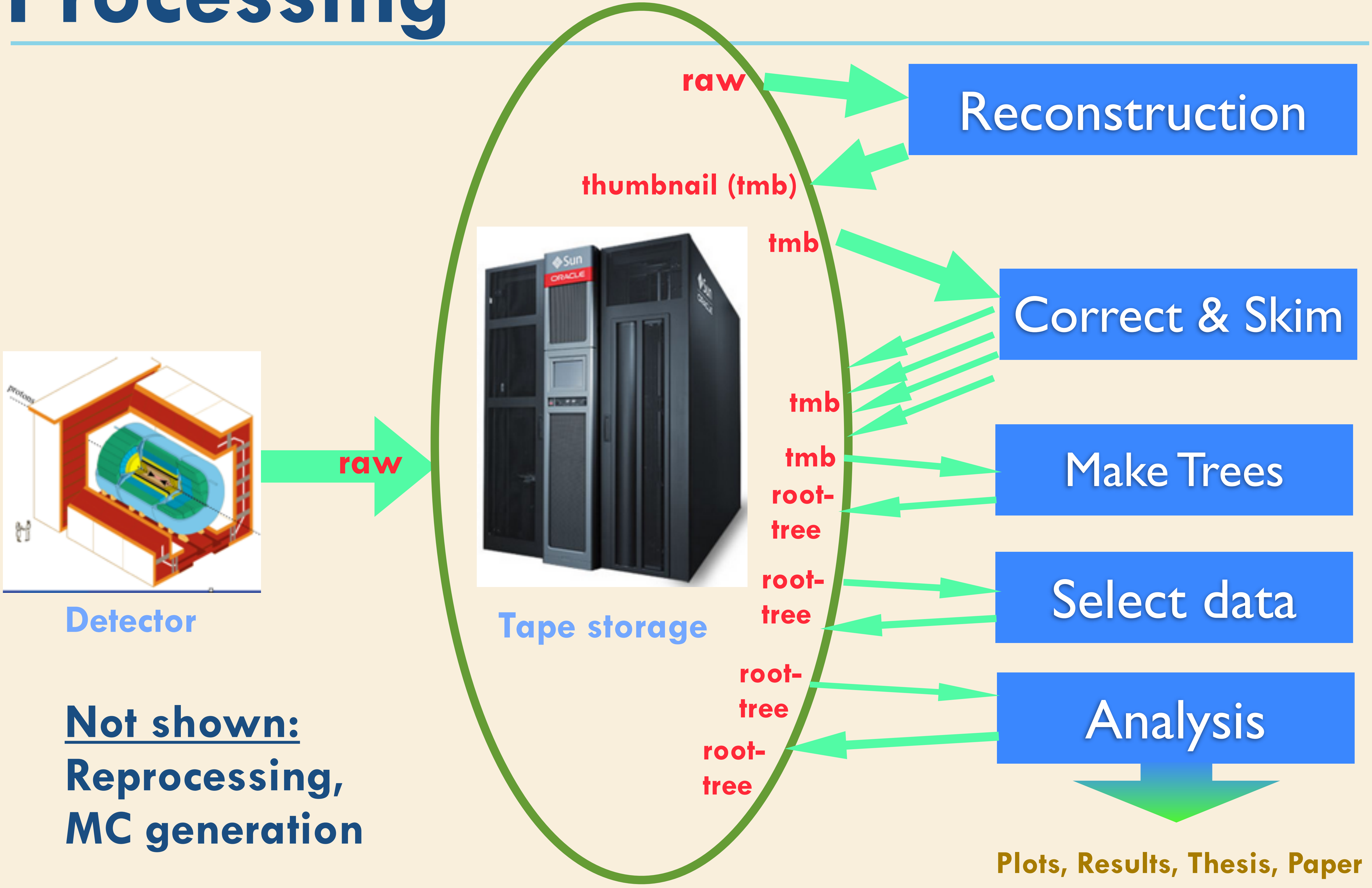
1999 – Detailed data management plan for D0 (SAM)

	Plan	Actual 2012
Entire dataset	0.5 - 1 PB	8.2 PB
Tapes	50 GB, 6 MB/s	800 GB, 120 MB/s
Cache	20 TB	760 TB
# files in catl	1M	140M
# events in catl	few billion	abandoned (scale issues)

150x Run I

Lesson: Think big and prepare for a lot bigger (scalability)

Processing



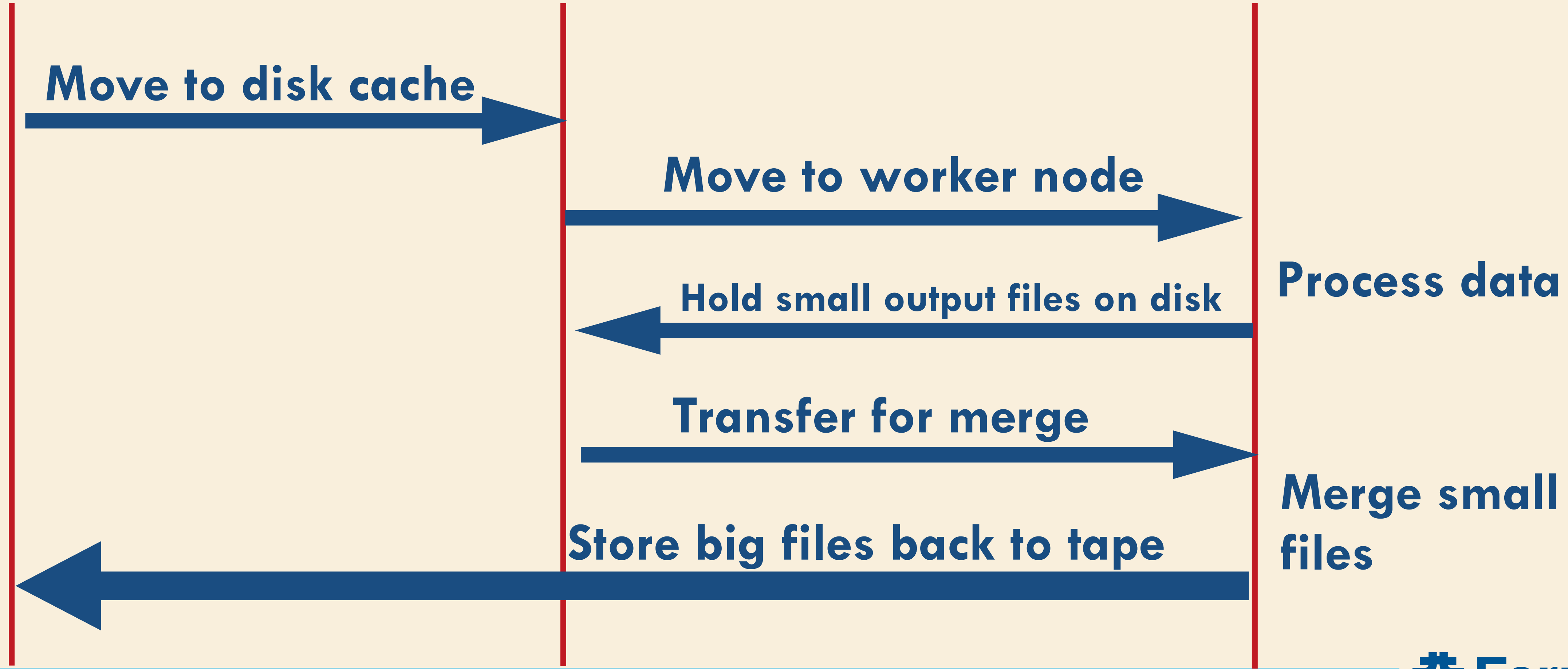
Not shown:
Reprocessing,
MC generation

Behind the scenes

Tape storage

Cache

Compute farm



Move to disk cache

Move to worker node

Process data

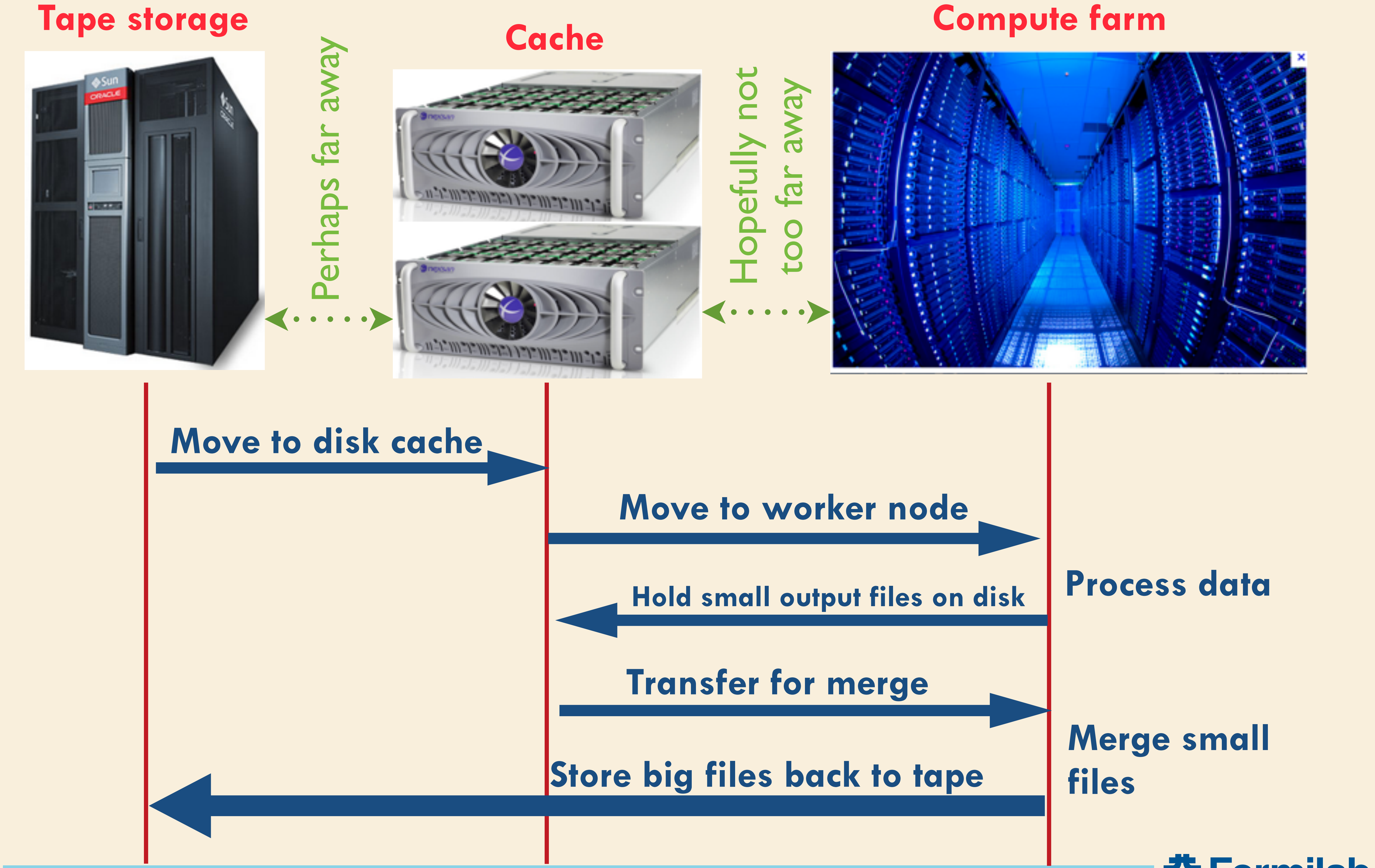
Hold small output files on disk

Transfer for merge

Merge small files

Store big files back to tape

Behind the scenes



Tape storage

Cache

Compute farm

Perhaps far away

Hopefully not too far away

Move to disk cache

Move to worker node

Process data

Hold small output files on disk

Transfer for merge

Merge small files

Store big files back to tape

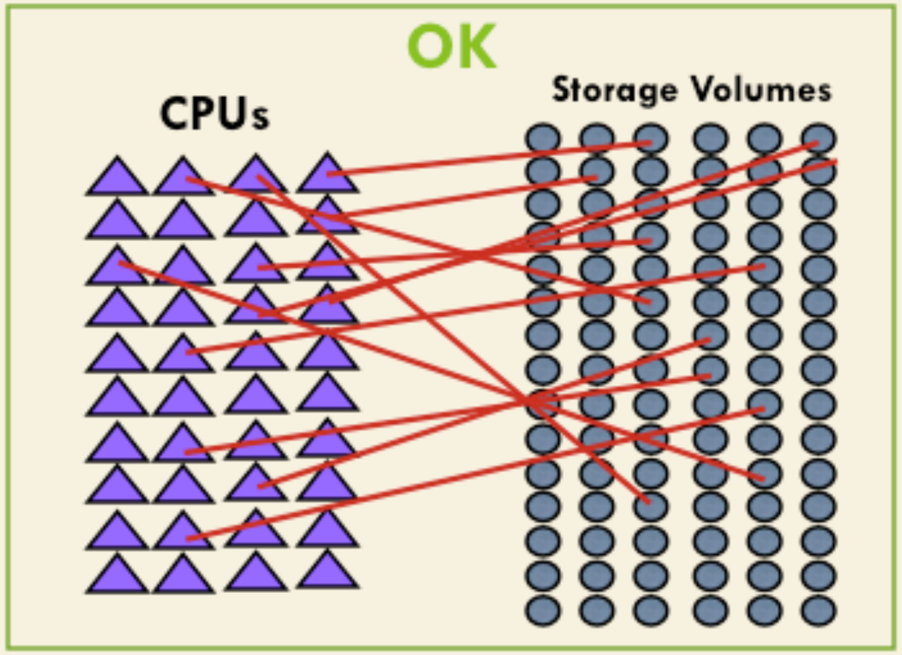
Sequential Access via Metadata (SAM)

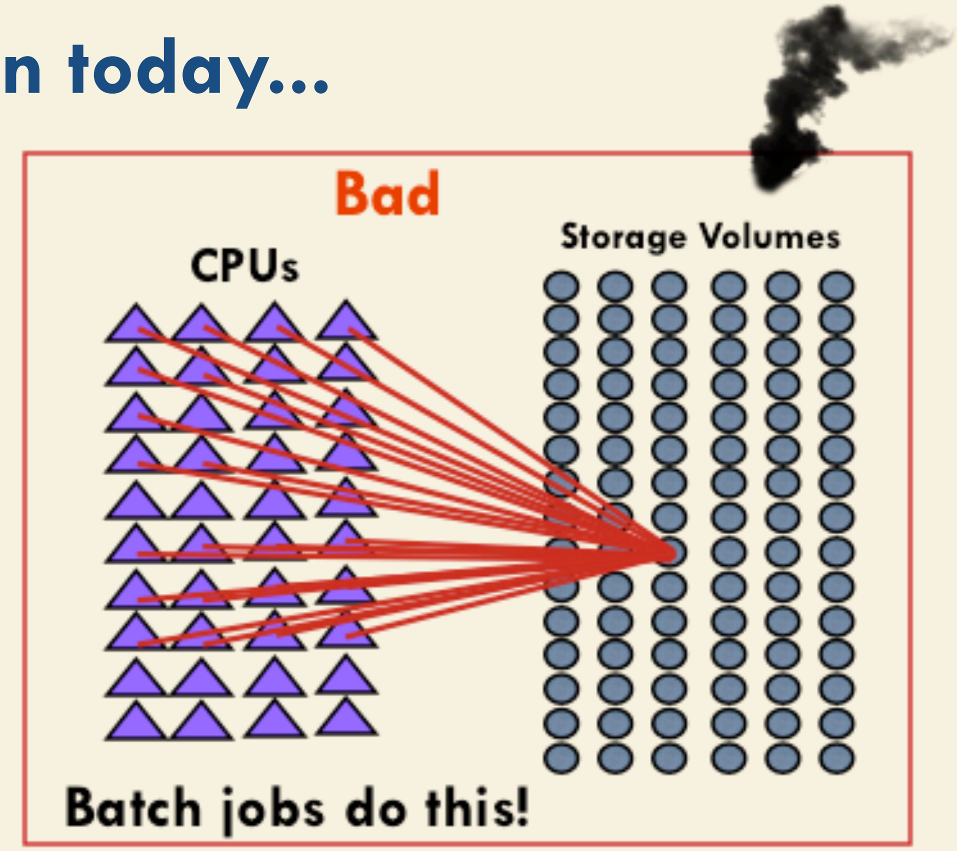
- A long evolution: 1998 – Present (16 years!)
  Started as a D0 project, then deployed to CDF in '03

- In early days storage systems were fragile,
  SAM layer provided needed throttling and protection
  reliability & scalability – continue to be important lessons

Aside: Storage systems can still be fragile even today…

Central NFS Storage





Batch jobs do this!

SAM for Tevatron Data Handling

- Forerunner of LHC data management (but important differences on coming slide)

- Metadata catalog (file name, size, events, run info, luminosity, MC details, …)
  Lesson: keep metadata out of the filename

- Dataset creation and queries
  Lesson: need to balance ease of use and flexibility (experts and non-experts)

- Replica catalog (where all file copies are located)

- Coordinates and manages data movement to jobs (bbftp, gridftp, dccp, SRM, future XRootD)
  Lesson: be ready to adopt new protocols - flexibility is key

- Cache management (now using dCache)
  Lesson: don't do it yourself once something more standard meets requirements

- File consumption and success tracking, recovery
  Lesson: this is an important feature (especially with opportunistic resources)

SAM on remote sites (proto-grid)

circa 2004 - Reprocessing

Early sites in US, Europe (Wuppertal, Prague), Brazil

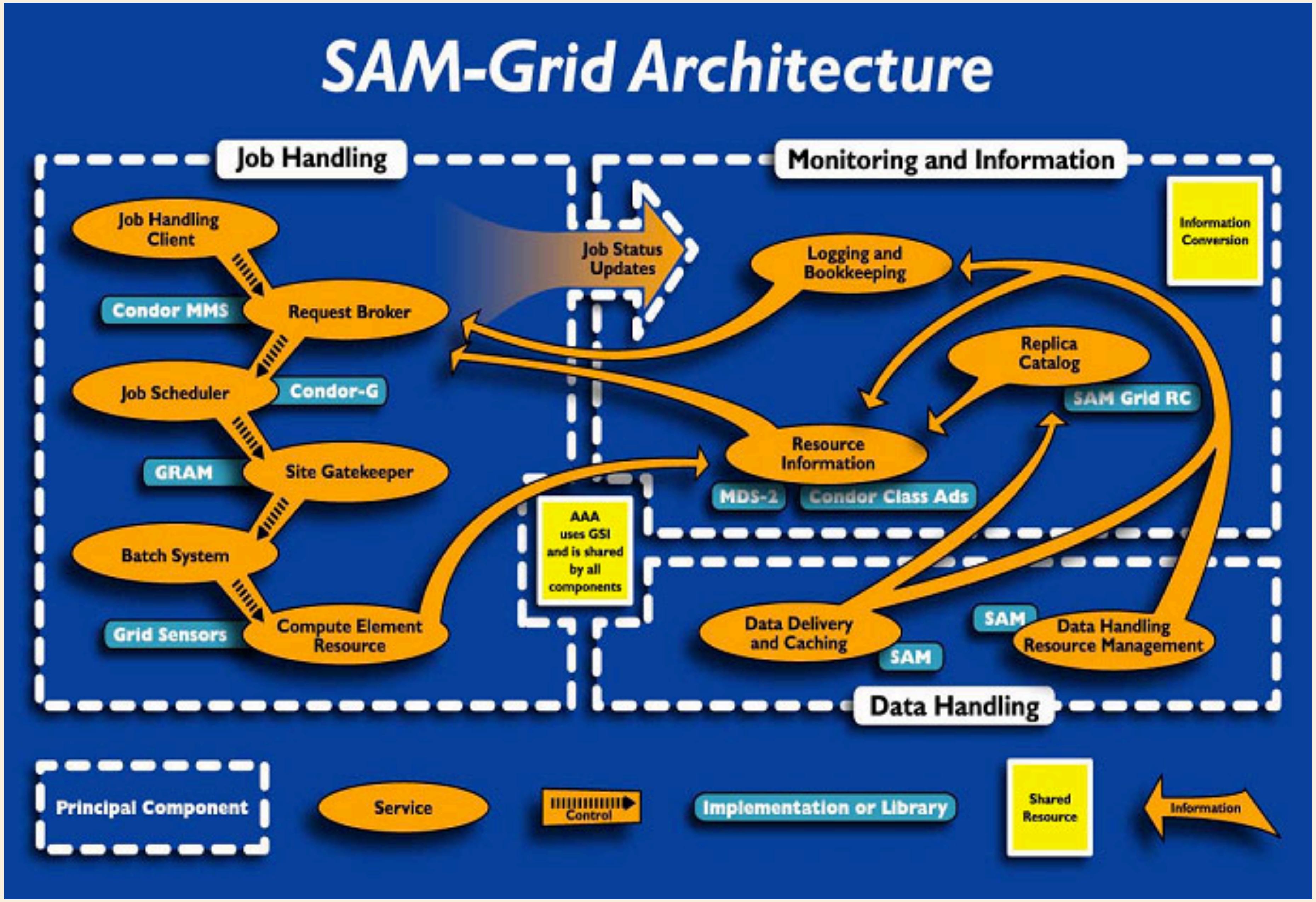
Wrote our own middleware while standard middleware was maturing

Required a large effort to maintain and operate

Lesson: Migrate to standard tools when they mature (much better when someone else maintains and operates). An important theme.

Condor, Globus, VDT, OSG

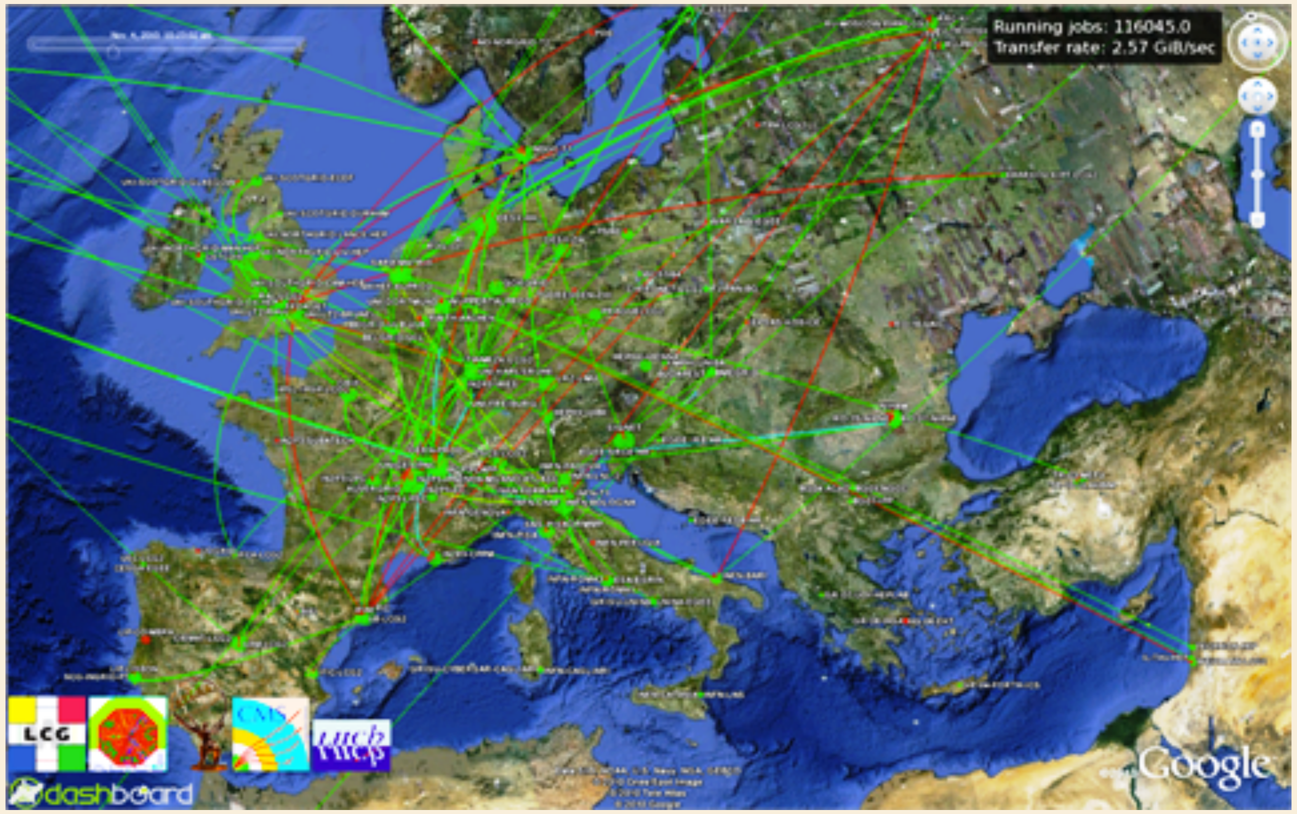
Lesson: Not all sites are the same - fast & slow queues



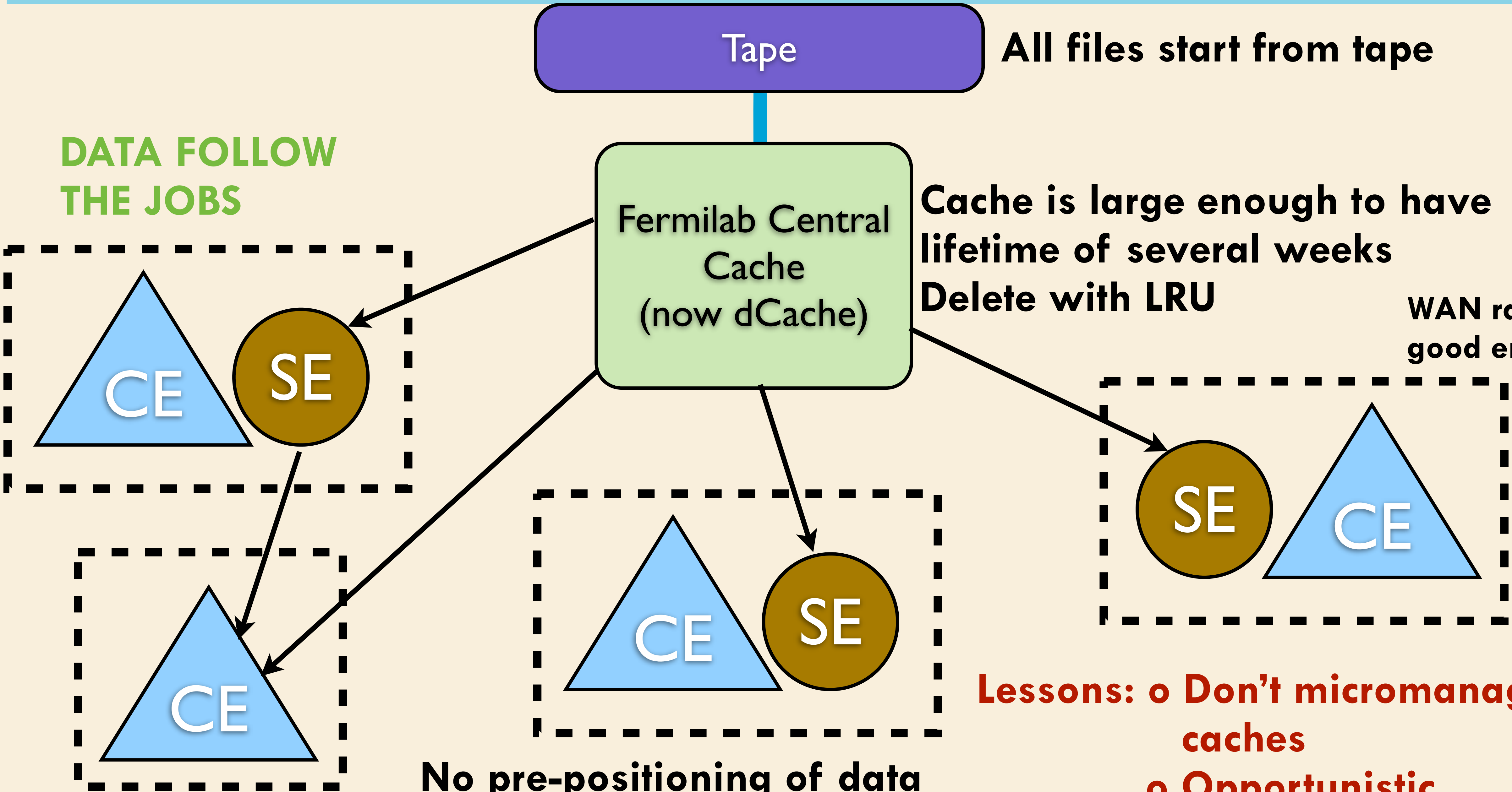
Grid Organizations

United States
Gulf of California
Mexico
Gulf of Mexico
Cuba
Caribbean
Guatemala
Nicaragua
Venezuela
Guyana
Suriname
Colombia
Ecuador
Peru
Brazil
Bolivia
Paraguay
North Atlantic Ocean

- 2007 submitting to OSG/LCG
- Support for remote sites is crucial:
  - Software stack (condor, globus, vdt)
  - Troubleshooting assistance
  - Infrastructure
  - Security
- LHC experiments also provide support



SAM Philosophies



All files start from tape

DATA FOLLOW THE JOBS

Cache is large enough to have lifetime of several weeks
Delete with LRU

WAN rates good enough

No pre-positioning of data beyond normal prestaging

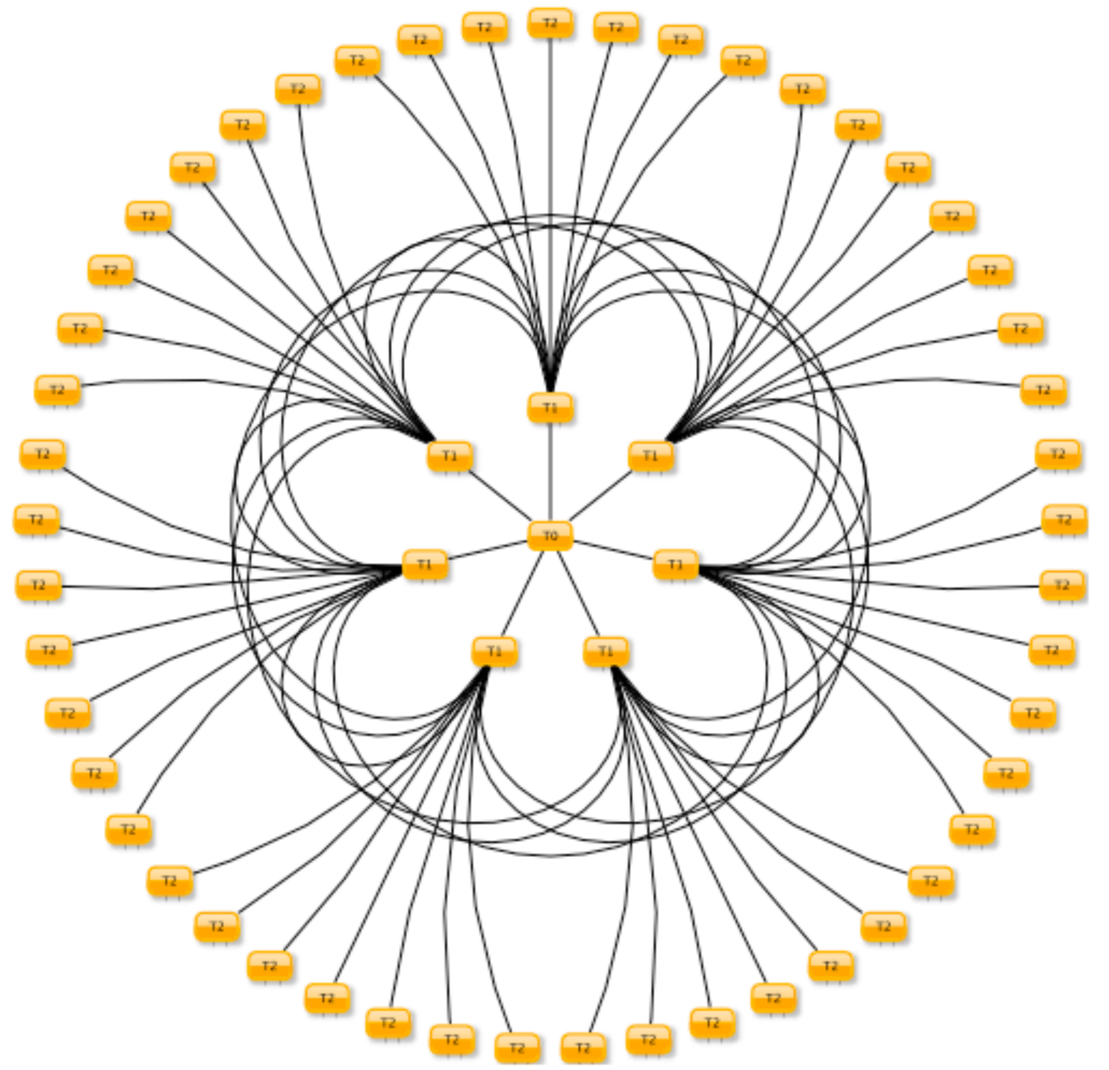
Lessons:
- Don't micromanage caches
- Opportunistic running requires agility

Favor **scalability** over **efficiency**

Run as many sites as we can, even on ones less productive

LHC Data Handling (CMS/ATLAS)

- Similar functionality to Tevatron Data Handling. The basic pieces are there, but implemented differently (PhEDEx, DBS3, PANDA, RUCIO, WMAgent/CRAB3, DIRAC)



- LHC Run 1: Hierarchical structure
  - Tier 0 @ CERN: Prompt reco
  - Tier 1 (regional): Re-reco, skimming calibration, make physics format
  - Tier 2 (local): analysis & MC

- Assign datasets to T1's and send jobs to the right place (JOBS FOLLOW THE DATA)

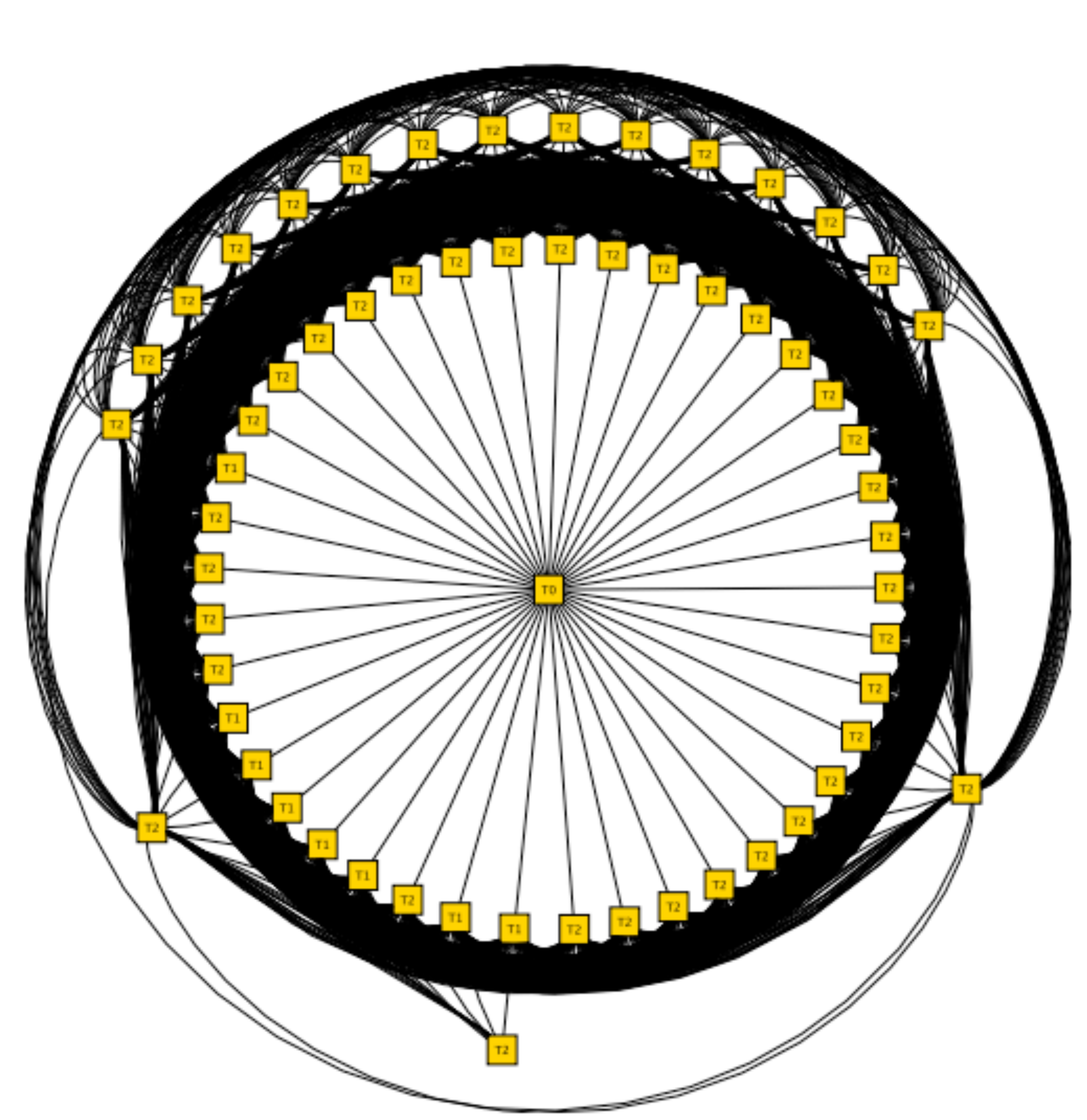
- Isolation of tape, limiting data movement allows for slow/non-robust networks, but no opportunistic running

Favor efficiency over scalability

We control a set of sites, so maximize their productivity

Future LHC Data Handling

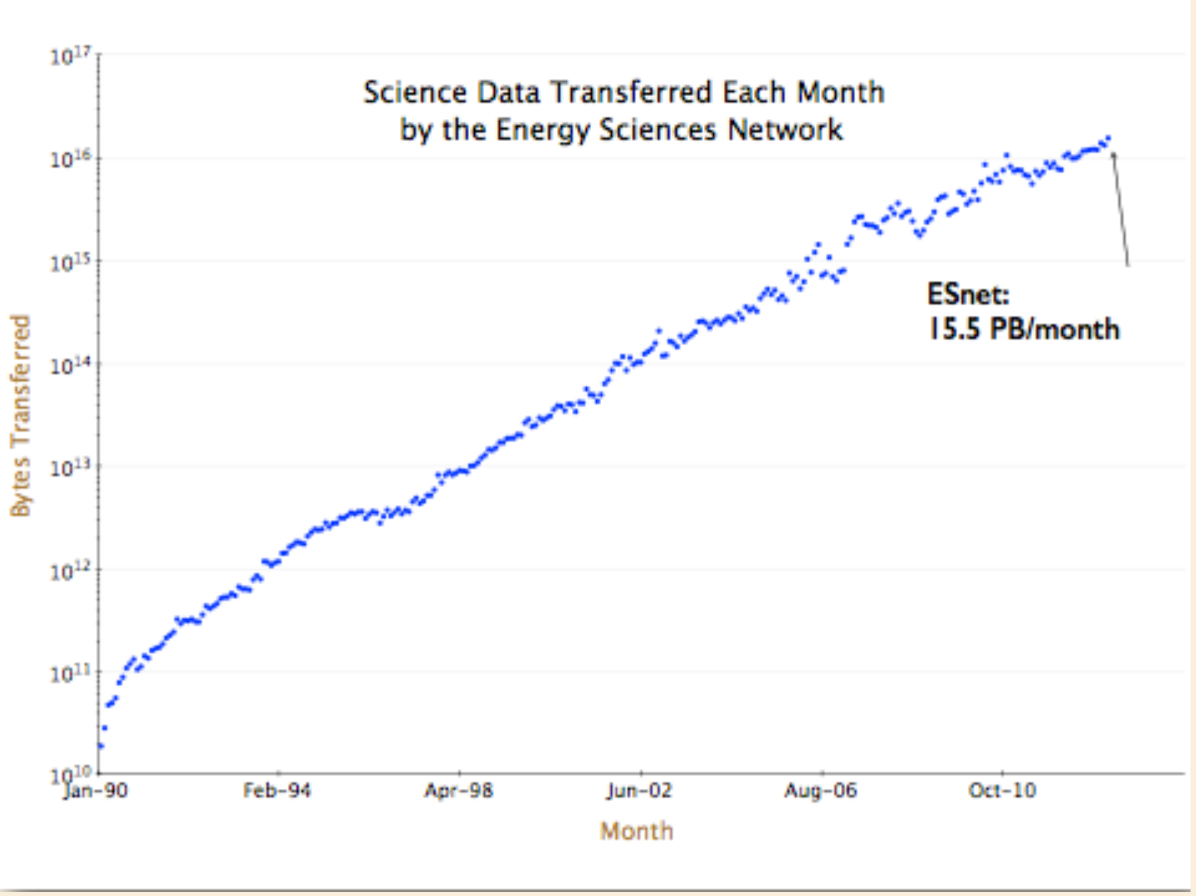
- Preparing for Run 2, the next level
- Networks are more robust and faster than predicted
- Connect everything together
- Still pre-position data on T1's but use federated XRootD as a fallback (Any data, Anytime, Anywhere) - mentioned many times here
- Makes opportunistic running possible (crucial for LHC Run 2)



Maximize reliability, scalability, and efficiency

Wow - the best of all worlds

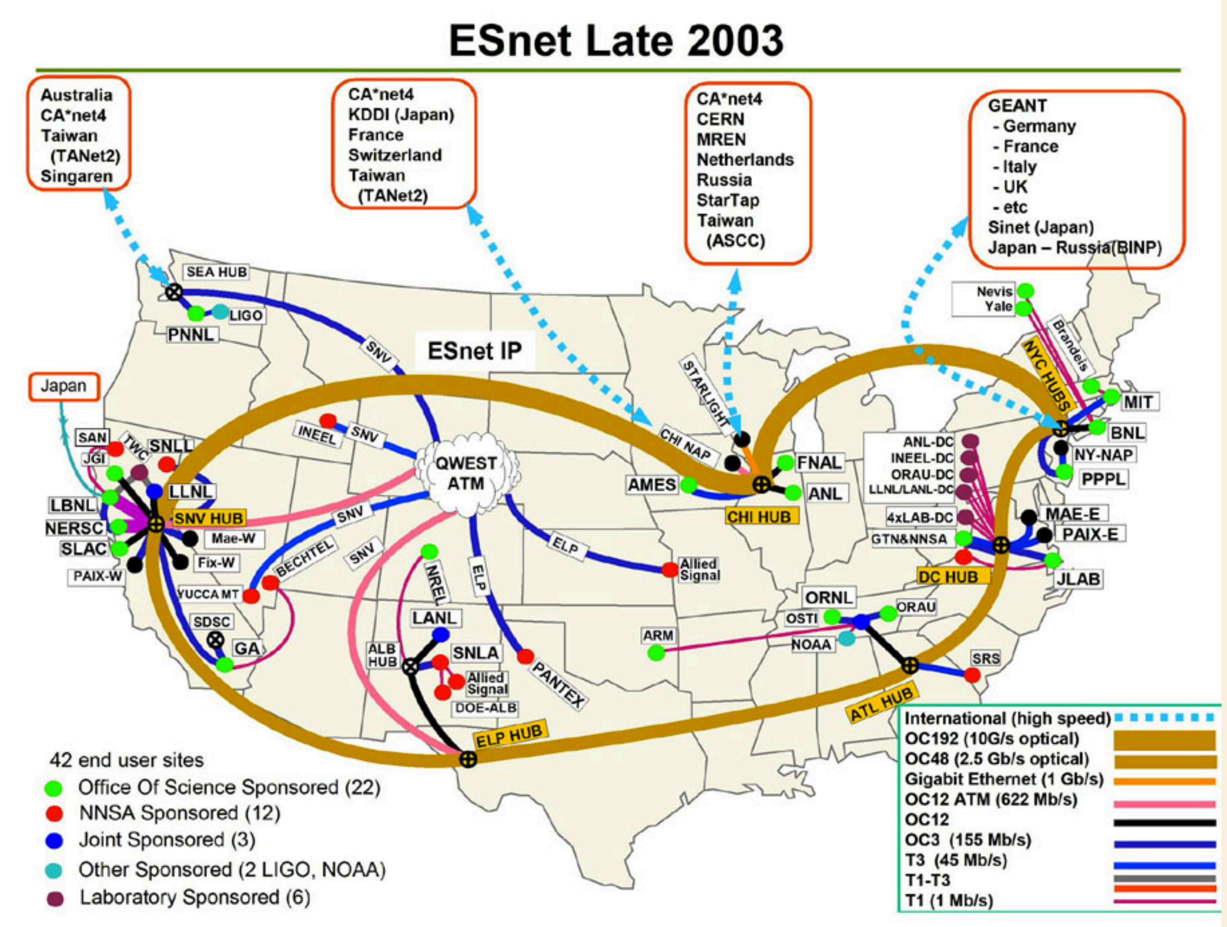
Robust Networks

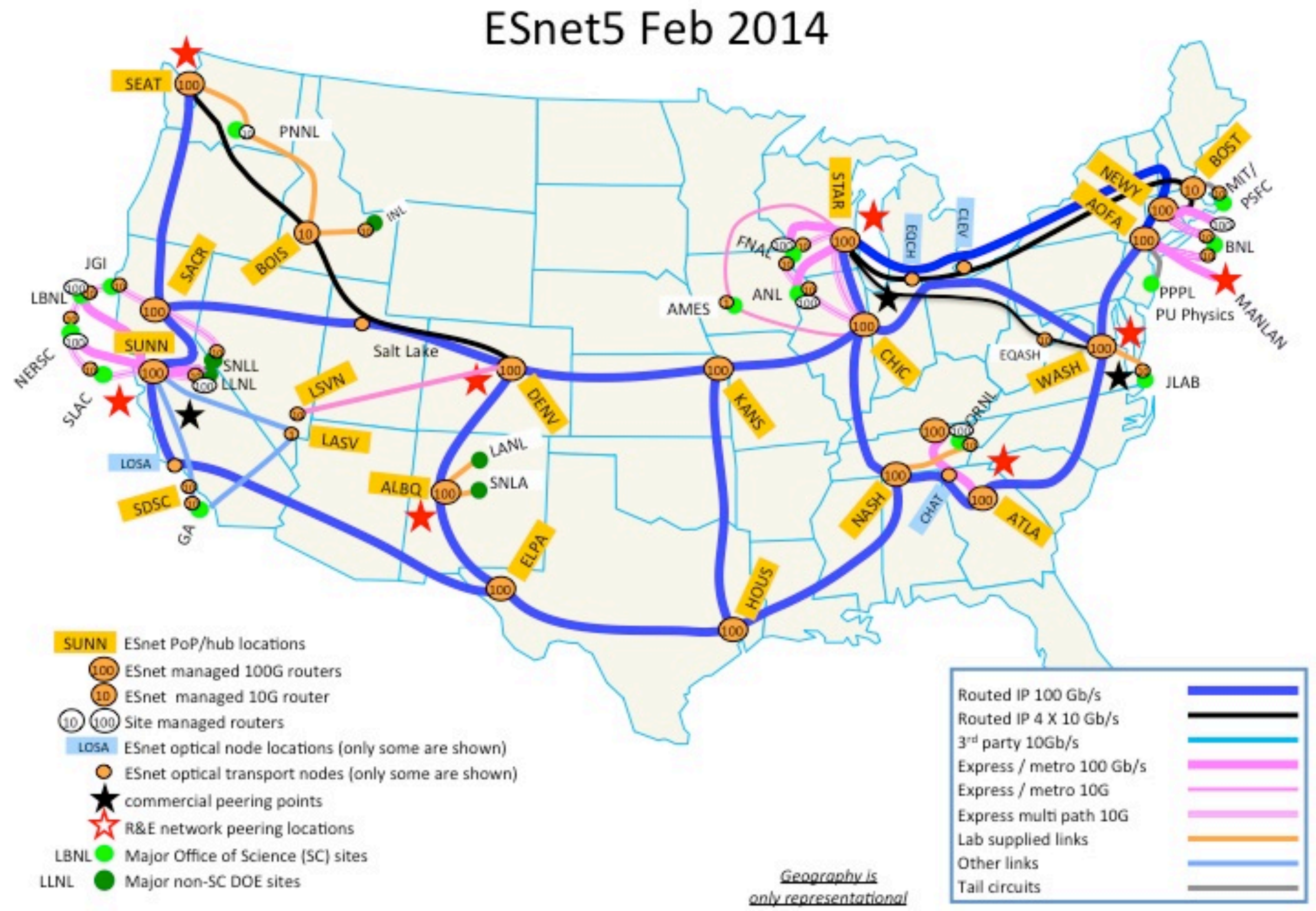


2003: 10 GB/s trunk lines
2014: 100 GB/s trunk lines

Large transcontinental pipes

ESNet From GB to PB


ESnet Late 2003


ESnet5 Feb 2014
SEAT
PNNL
INL
BOIS
SACR
JGI
LBNL
NERSC
SUNN
SNLL
LLNL
SLAC
Salt Lake
LSVN
LASV
LOSA
SDSC
GA
DENV
LANL
SNLA
ALBQ
ELPA
KANS
HOUS
AMES
FNAL
ANL
STAR
EQCH
CLEV
CHIC
NASH
CHAT
ORNL
ATLA
EQASH
WASH
JLAB
NEWY
AOFA
BOST
MIT/PSFC
BNL
PPPL
PU Physics
MANLAN
SUNN ESnet PoP/hub locations
100 ESnet managed 100G routers
10 ESnet managed 10G router
10 100 Site managed routers
LOSA ESnet optical node locations (only some are shown)
ESnet optical transport nodes (only some are shown)
★ commercial peering points
☆ R&E network peering locations
LBNL ● Major Office of Science (SC) sites
LLNL ● Major non-SC DOE sites
Geography is only representational
Routed IP 100 Gb/s
Routed IP 4 X 10 Gb/s
3rd party 10Gb/s
Express / metro 100 Gb/s
Express / metro 10G
Express multi path 10G
Lab supplied links
Other links
Tail circuits

Near future: Use new resources

- Leadership class HPC machines Make look like an OSG site with OSG-Connect and Parrot

- Take advantage of coming Exascale computing?

- Cloud resources to provision worker nodes on demand for peak activity ("Grid Bursting")

But are they suitable for HEP jobs (can be long, high i/o)? Cost effective?

See Garzoglio talk about "On Demand" resources

The Future

- The computing landscape is under great change
- CPUs can't go faster (no 10 GHz?)
  - Limited by power density
  - So integrate more of them together
  - Multicore CPUs
  - GPGPUs (with attached ARMs)
  - Intel PHI co-processors - Many integrated cores (MICs)
- Using these technologies is challenging, especially for non-experts
- How are data fed to these devices with reliability, scalability, and efficiency?

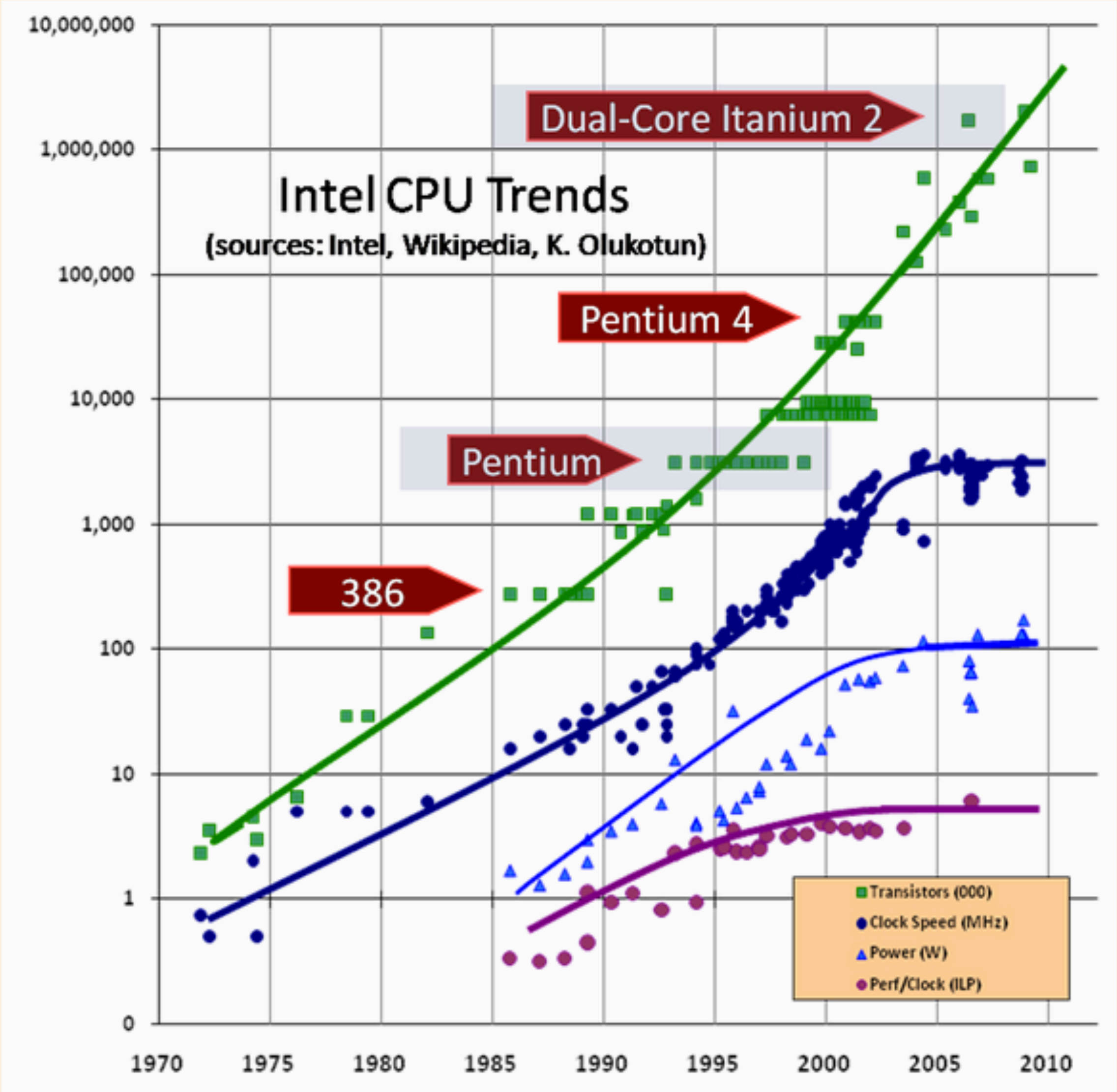
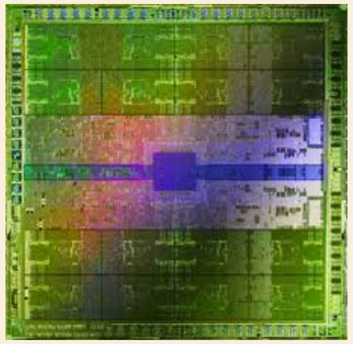
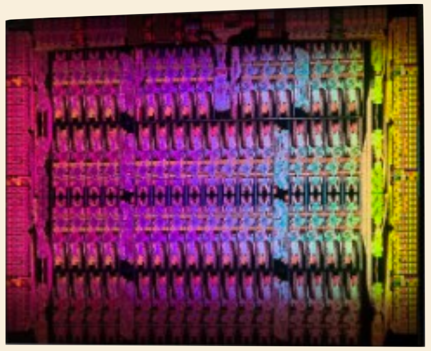


Figure 1: Intel CPU Introductions (graph updated August 2009; article text original from December 2004)





Disruptive technology?

- o Does industry's vision of Big Data (Map-Reduce/Hadoop) have a place in HEP?

- o Will we need non-traditional High Throughput Computing (HTC) machines (Data Appliances)?

- o Does the notion of a file survive? E.g. No-SQL databases for events? But remember the lessons from Objectivity!!

- o Can we use Data Appliances?
  e.g. EMC Isilon - Very high speed i/o with fast disk caching - can perform map-reduce & DB functions in the cache! Disks become compute nodes! Wow!

  e.g. YARC - "purpose built" data appliance from Cray
  Does a "graph analysis" to find connections in data

YarcData
Getting to Eureka! faster

- o Must pre-load these appliances. Is opportunistic use possible?
  How would we use these with reliability, scalability, and efficiency?

Is a Paradigm Shift coming?

- Collider experiments (CMS, Atlas, …) write event based data. Can this model hold up for the far future? What other models are there?

- Does the event model work for the next generation neutrino and muon experiments with trigger-less online? Are they more digital signal processing? Do we need a new set of data management and analysis tools? Just starting…

- Collider experiments involve analyzing a subset of the full data (skims). But other experiments may need to analyze EVERYTHING. What do we need to make that work?

- What will it take to reach the next level?

Summary

- There's no time for rest! New experiments & runs are coming and they must succeed!
- Particle physics has an excellent history of grasping the next level of data management to reach the next level of discovery
- Ability and willingness to adjust when reliability, scalability, and efficiency can improve is the key and will allow us to reach that next level
- Are there new machines, technologies, paradigm shifts that take us to the next level with more capabilities than before?